PROVABLY EFFICIENT EXPLORATION FOR OFF-DYNAMIC RESTLESS MULTI-ARMED BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online learning in Restless Multi-Armed Bandits (RMABs) under distributional shift: the agent has access to offline data from a source environment whose dynamics differ from the target by an unknown, arm-specific gap. Ignoring this data wastes a valuable signal and incurs high initial regret; blindly trusting it risks catastrophic failure when the gap is large. We propose `ODR-IOB`, the first provably safe and adaptive framework for off-dynamic RMABs, built around a *Gap-Aware Intersection-of-Balls* confidence set construction. The key idea is to maintain a pure-online ball as a safety net, and an offline-augmented ball whose size adapts to a data-driven estimate of the dynamics gap, and plan optimistically over their intersection. The framework is provably never worse than pure online learning, regardless of how large or misleading the gap turns out to be. When the gap is small and the source data is informative, the intersection contracts strictly, yielding a tighter optimistic model and lower regret. We formalize these properties as three simultaneous guarantees: the true dynamics always lie inside the intersection with high probability (*coverage*); performance never degrades below the pure-online baseline (*safety*); and regret improves proportionally to the fraction of the state-action space unaffected by the shift (*acceleration*). Experiments confirm that `ODR-IOB` significantly outperforms the pure online baseline under small gaps while gracefully degrading to it under large gaps.

1 INTRODUCTION

Sequential decision-making under uncertainty is central to many high-stakes real-world applications, from public health interventions to resource allocation in constrained systems. Restless Multi-Armed Bandits (RMABs) (Whittle, 1988) provide a principled framework for these problems: an agent activates a limited budget of arms at each step, each arm evolving according to its own unknown Markov dynamics, and the goal is to maximize cumulative reward over time. RMABs have been deployed across a diverse range of domains, including public health (Mate et al., 2022), anti-poaching patrol (Killian et al., 2023), and wireless scheduling (Xiong et al., 2022b). In such domains, the standard approach of learning purely from online interaction is untenable in practice. Online algorithms (Ortner et al., 2012; Wang et al., 2020; Jung & Tewari, 2019; Xiong et al., 2022a; Wang et al., 2023; Xiong & Li, 2023; Akbarzadeh & Mahajan, 2022; Xiong & Li, 2024) must explore, i.e., activating arms to estimate their dynamics, and this exploration phase directly translates to suboptimal decisions during a period when the cost is highest. In healthcare (Mate et al., 2022), a mother who disengages during exploration may be lost to follow-up entirely; in infrastructure management, a suboptimal decision during early deployment can have cascading consequences. The regret incurred during online exploration is not merely a computational inconvenience, but represents real harm.

Practitioners deploying an RMAB in a new environment (e.g., a rural community) typically possess a large offline dataset from a related but different source environment (Mate et al., 2022; Kong et al., 2026) (e.g., historical data from an urban deployment). This dataset encodes valuable structural information about arm dynamics and could, in principle, dramatically reduce the exploration cost in the new setting. The temptation is to simply trust this data and deploy a policy optimized for the source environment. However, the source and target environments differ, and the dynamics gap between them is unknown, heterogeneous across arms, and potentially large. Blindly trusting offline data from a mismatched source can lead to a policy that is persistently suboptimal, accumulating

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

regret without bound (Eysenbach et al., 2021; Niu et al., 2022; Qu et al., 2024) as it keeps making the same wrong decisions with misplaced confidence.

This sets up a fundamental dilemma. Two natural extremes both fail: *pure online learning* ignores the offline data entirely (Ortner et al., 2012; Xiong et al., 2022a; Wang et al., 2023), suffering from high initial regret that is unacceptable in health settings; *pure offline deployment* blindly trusts the source dynamics, risking catastrophic and persistent failure whenever the dynamics gap is non-trivial. The central challenge is what we call *Safe Acceleration*: leveraging offline source data to reduce initial regret, while guaranteeing that performance is provably never worse than learning from scratch. Crucially, this guarantee must hold without any prior knowledge of the dynamics gap, which is unknown and may be arbitrarily large.

Prior work on transfer in reinforcement learning has largely sidestepped this challenge. Empirical cross-domain methods (Eysenbach et al., 2021; Liu et al., 2022; Niu et al., 2022) show practical gains from transfer but offer no safety guarantees: they can silently fail when the source is misleading. On the theoretical side, Qu et al. (2024) establish a fundamental hardness result: without any prior information on the dynamics gap, source data *cannot* reduce sample complexity in the worst case. Their positive result is that the HySRL algorithm achieves efficiency gains only when a lower bound β on the gap is known in advance, a strong assumption that is often unavailable in practice. Hybrid RL methods (Xie et al., 2021; Song et al., 2023) exploit offline data from the *same* environment, bypassing the shift problem entirely. Distributionally robust approaches (Shi et al., 2023; Panaganti & Kalathil, 2022; Shi et al., 2024) handle worst-case shift but are inherently conservative: they optimize for the worst possible source and cannot exploit a well-matched source to accelerate learning. None of these approaches provide the combination we need: adaptive use of offline data, provable safety under arbitrary unknown shift, and strict efficiency gains when the shift is small.

The key insight is to *decouple safety from efficiency* by maintaining two confidence sets simultaneously, rather than constructing a single blended set. A natural first attempt is to form a weighted average of the offline and online kernels and build a single confidence set around this blend. This is fragile: the radius of such a set must account for the unknown dynamics gap, and underestimating the gap causes the set to exclude the true dynamics entirely, leading to vacuously optimistic planning and unbounded regret. Instead, we maintain a **pure-online ball** built solely from online interaction data. By construction, this ball always contains the true dynamics with high probability, regardless of the source environment. It serves as an unconditional safety net. An **offline-augmented ball** centered at a gap-weighted blend of the source and online kernels, with a radius that adapts to a data-driven upper confidence bound on the Wasserstein dynamics gap. This ball is tighter when the source is reliable and wider when it is not, serving as an adaptive accelerator.

Policy optimization is performed over the *intersection* of these two sets via an Extended Linear Program (ELP) over occupancy measures. The intersection is never wider than the online ball, guaranteeing safety by construction. When the gap is small, the intersection is strictly narrower, concentrating the feasible set around a more accurate model and enabling faster learning. Crucially, although the two balls are centered at different points, this offset does not spoil the shrinkage: once the offline radius is smaller than the online radius, the intersection width is governed by the smaller radius alone. We summarize the contributions as follows:

▷ We introduce the *Off-Dynamic RMAB* setting, formalizing online RMAB learning with offline data from a source environment with unknown, heterogeneous dynamics gaps (Section 3).

▷ We propose `ODR-IOB`, a five-step pipeline that estimates the Wasserstein gap, converts it to an ℓ_1 bias bound, blends source and online kernels with a gap-decaying weight, intersects the resulting confidence balls, and solves an ELP over the intersection (Section 4).

▷ We prove four nested guarantees: the gap estimator never underestimates the true gap; the true dynamics always lie in the IOB set; `ODR-IOB` is never worse than pure online; and the IOB set contracts strictly when the gap is small. Together these yield a regret bound that matches pure online learning in the worst case and improves by a factor of $\sqrt{|\mathcal{B}|/SA}$ otherwise (Section 5).

2 Related work

Related Work. Online RMAB learning (Whittle, 1988; Ortner et al., 2012; Xiong et al., 2022a; Wang et al., 2023; Xiong & Li, 2024) assumes the agent learns purely from target-environment interaction, offering no mechanism to exploit offline source data. Empirical transfer RL methods (Ey-

Table 1: Comparison of ODR-IOB with related frameworks. "Shift" indicates a dynamics gap between source and target domains. "Safety" indicates that the performance is not worse than pure online learning. "No Prior" indicates no prior knowledge of the gap magnitude. ✓: satisfied; ✗: not satisfied.

Category	Paper	Offline Data	Shift	Safety	No Prior	RMAB
Online RMAB	Xiong et al. (2022a); Wang et al. (2023)	✗	✗	✓	✓	✓
Hybrid RL	Xie et al. (2021); Song et al. (2023)	✓	✗	✓	✓	✗
Transfer RL	Eysenbach et al. (2021); Niu et al. (2022)	✓	✓	✗	✓	✗
Transfer RL (theory)	Qu et al. (2024)	✓	✓	✓	✗	✗
Robust RL	Shi et al. (2023); Panaganti & Kalathil (2022)	✓	✓	✓	✓	✗
This Work (ODR-IOB)	—	✓	✓	✓	✓	✓

senbach et al., 2021; Liu et al., 2022; Niu et al., 2022; 2023) show practical cross-domain gains but provide no formal safety guarantees and can silently fail under large dynamics gaps. Qu et al. (2024) establish a fundamental hardness result. Without prior knowledge of the gap, source data cannot reduce sample complexity, and their positive result HySRL requires a known gap lower bound β and a separate reward-free identification phase. Hybrid RL (Xie et al., 2021; Song et al., 2023) exploits offline data from the *same* environment, bypassing the shift problem entirely. Distributionally robust RL (Shi et al., 2023; Panaganti & Kalathil, 2022) handles worst-case shift but cannot exploit a well-matched source to accelerate learning. ODR-IOB is, to our knowledge, the first framework to simultaneously achieve online learning in the RMAB setting, safe exploitation of offline data with *unknown* dynamics shift, and provable acceleration when the shift is small-all without any prior gap knowledge; see Table 1 for reference.

3 Off-Dynamic RMAB

We consider an RMAB setting with N arms. The underlying process of each arm $n \in \{1, \dots, N\}$ is modeled as a Markov Decision Process (MDP), denoted by $\mathcal{M}_n = (\mathcal{S}, \mathcal{A}, R, P_n)$. In particular, $\mathcal{S}$ is a finite state space with $|\mathcal{S}| = S$, $\mathcal{A} = \{0, 1\}$ is the binary action space (passive/active), and $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the known predefined reward function. Notice that $P_n : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the *unknown* state transition probability kernel for arm n, and arms differ from each other with this unknown transition kernel. At each time step t, a policy π_t selects a subset of K arms to activate, where $K < N$.

3.1 Offline Source Data and Source Prior

We have access to a static offline dataset $\mathcal{D}_{\text{src}}$ collected from a population of arms with diverse dynamics: $\mathcal{D}_{\text{src}} = \{(s_{n,i}, a_{n,i}, s'_{n,i})_i^{M_n}\}_{n=1}^N$, where M_n corresponds to the total samples of the distinct arm (implicitly characterized by its feature) n. Different arms may exhibit heterogeneous transition behaviors, but all data are drawn from the same source environment. For each arm n, we estimate an empirical transition kernel from its offline samples as

$$\widehat{P}_n^{\text{src}}(\cdot \mid s, a) = \frac{\sum_i \mathbb{1}\{s'_{n,i} = \cdot, s_{n,i} = s, a_{n,i} = a\}}{\sum_i \mathbb{1}\{s_{n,i} = s, a_{n,i} = a\}}. \tag{1}$$

The corresponding *source prior model* for arm n is thus $P_n^{\text{src}}(\cdot \mid s, a) := \widehat{P}_n^{\text{src}}(\cdot \mid s, a)$, which serves as the initial value of the arm's dynamics before any online interaction.

Remark 3.1. Throughout the paper we treat P_n^{src} as exact, i.e., we assume M_n is large enough that estimation error in the source kernel is negligible. Source-side uncertainty can nevertheless be incorporated into the IOB framework accordingly.

3.2 The Dynamics Gap

The source kernel P_n^{src} is an imperfect proxy for the true target dynamics $P_n^\star$. We quantify their discrepancy via the *dynamics gap*.

Definition 3.2 (Dynamics Gap)**.** For arm n and state-action pair (s, a), the *dynamics gap* is

$$\Delta_n(s, a) \;=\; W_1^\rho\big(P_n^{\text{src}}(\cdot \mid s, a),\; P_n^\star(\cdot \mid s, a)\big), \tag{2}$$

where W_1^ρ is the Wasserstein-1 distance under ground metric $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$. The gap $\Delta_n(s,a)$ is *unknown* to the agent and may be heterogeneous across arms and state-action pairs.

Assumption 3.3 (Ground Metric)**.** The metric ρ satisfies $d_{\min} \triangleq \min_{x \neq y} \rho(x,y) > 0$ and $d_{\max} \triangleq \max_{x,y} \rho(x,y) < \infty$.

3.3 Objective: Regret Minimization

Let π^* be the optimal policy that has full knowledge of the true dynamics $\{P_n\}_{n=1}^N$. Let π_t be the policy deployed by the learning algorithm at time t. The objective is to design an algorithm that generates a sequence of policies $\{\pi_t\}$ to minimize the cumulative regret over a time horizon T. The regret $\mathcal{R}_T$ is defined as the difference in expected total rewards between the optimal policy and the learning algorithm's policy: $\mathcal{R}_T = \sum_{t=1}^T \left(\mathbb{E}[r(t)|\pi^*] - \mathbb{E}[r(t)|\pi_t]\right)$, where $r(t) := \sum_n R_n(s_n(t), a_n(t))$ is the total reward obtained at time t. The central research question is to design an online algorithm that intelligently estimates the dynamics gaps $\{\Delta_n(s,a)\}$ and uses these estimates to modulate its reliance on the priors $\{P_n^{src}\}$, in order to achieve a regret $\mathcal{R}_T$ that grows sub-linearly in T.

4 Gap-Aware ODR−IOB Framework

The central challenge is to fuse offline source knowledge with online observations in a provably safe way. Blending the two models with a weighted average is only as safe as the weight chosen, and choosing correctly requires knowing precisely how much the source dynamics differ from the target. Underestimate that gap and the confidence set may exclude the true dynamics entirely, producing catastrophically unbounded regret. Our solution is to decouple safety from efficiency. ODR−IOB maintains two confidence sets in parallel: a pure-online ball that always contains the true dynamics and serves as an unconditional safety net, and an offline-augmented ball centered at a gap-weighted blend of the two models, acting as an adaptive accelerator when source data is reliable. Policy optimization is performed over their intersection. Because the intersection is always contained within the online ball, it is never wider than the pure online baseline, guaranteeing sound performance regardless of how misleading the source turns out to be. When the gap is small, the intersection becomes strictly narrower, yielding a sharper optimistic policy and lower regret. In the following, we introduce the detailed procedures of ODR−IOB shown in Figure 1.

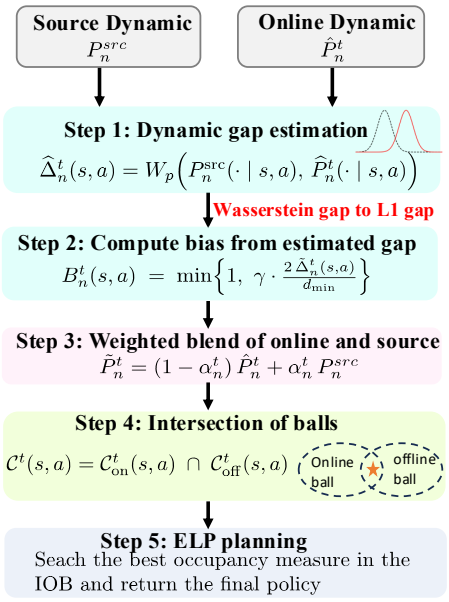

Figure 1: The proposed ODR−IOB pipeline.

4.1 Step 1: Dynamics Gap Estimation

For each arm n and state–action pair (s,a), we quantify the distributional discrepancy between the source prior and the online empirical transition kernel through a **dynamics gap**:

$$\widehat{\Delta}_n^t(s,a) = W_1^\rho\left(P_n^{\text{src}}(\cdot \mid s,a),\ \widehat{P}_n^t(\cdot \mid s,a)\right), \tag{3}$$

where W_1^ρ denotes the discrete Wasserstein distance defined over the ground metric $\rho(\cdot,\cdot)$ on the state space $\mathcal{S}$. For arm n and state-action pair (s,a), let $C_n^t(s,a)$ denote the number of observed transitions from state-action pair (s,a). The corresponding empirical transition kernel is computed as $\widehat{P}_n^t(\cdot \mid s,a) = \frac{C_n^t(s,a;\cdot)}{\max\{1, C_n^t(s,a)\}}$. For discrete $\mathcal{S}$, equation 3 reduces to a linear program. Setting

$\mu = P_n^{\mathrm{src}}(\cdot \mid s,a)$, $\nu = \hat{P}_n^t(\cdot \mid s,a)$, and cost matrix $C_{ij} = \rho(s_i, s_j)$:

$$\widehat{\Delta}_n^t(s,a) = \min_{\pi \geq 0} \sum_{i,j} C_{ij}\, \pi_{ij} \quad \text{s.t.} \quad \textstyle\sum_j \pi_{ij} = \mu_i, \quad \sum_i \pi_{ij} = \nu_j. \tag{4}$$

The transport plan π_{ij} moves probability mass from s_i under the source to s_j under the online kernel, and the objective is the minimum total cost of this redistribution. When $|\mathcal{S}|$ is small, equation 26 is solved exactly; for larger state spaces, entropic regularization (e.g., Sinkhorn) yields a negligible approximation error.

High-probability correction. The empirical gap $\hat{\Delta}_n^t$ compares P_n^{src} against $\hat{P}_n^t$, not against the unknown true kernel $P_n^\star$. We therefore correct for the finite-sample error in $\hat{P}_n^t$.

Lemma 4.1. *Let $\epsilon_n^t(s,a)$ denote a high-probability deviation bound between the empirical transition $\widehat{P}_n^t$ and the true underlying dynamics $P_n^\star$. By the multinomial Weissman inequality, with probability at least $1-\xi$:*

$$\|\hat{P}_n^t(\cdot \mid s,a) - P_n^\star(\cdot \mid s,a)\|_1 \leq \epsilon_n^t(s,a) \;\triangleq\; \sqrt{\frac{2\ln\big((2^{|\mathcal{S}|}-2)/\xi\big)}{\max\{1,\, C_n^t(s,a)\}}}. \tag{5}$$

To lift this ℓ_1 bound to the Wasserstein domain, we use the TV–W_1 sandwich inequality (Lemma E.1):

$$W_1^\rho(\hat{P}_n^t, P_n^\star) \leq \frac{d_{\max}}{2}\, \|\hat{P}_n^t - P_n^\star\|_1 \leq \frac{d_{\max}}{2}\, \epsilon_n^t(s,a) \;\triangleq\; \xi_n^t(s,a). \tag{6}$$

Confidence-corrected gap. By the triangle inequality for W_1^ρ:

$$W_1^\rho(P_n^{\mathrm{src}}, P_n^\star) \leq \underbrace{W_1^\rho(P_n^{\mathrm{src}}, \hat{P}_n^t)}_{=\hat{\Delta}_n^t} + \underbrace{W_1^\rho(\hat{P}_n^t, P_n^\star)}_{\leq \xi_n^t}. \tag{7}$$

We therefore define the *confidence-corrected gap*:

$$\tilde{\Delta}_n^t(s,a) = \hat{\Delta}_n^t(s,a) + \xi_n^t(s,a). \tag{8}$$

By construction, $\tilde{\Delta}_n^t$ is a high-probability upper confidence bound on the true dynamics gap $\Delta_n(s,a) = W_1^\rho(P_n^{\mathrm{src}}, P_n^\star)$ (formalized in Theorem 5.1). It serves as the key input to Steps 2–4: the larger $\tilde{\Delta}_n^t$, the less the algorithm trusts P_n^{src}, and the wider the offline confidence ball becomes, automatically preventing over-reliance on a misleading source.

4.2 Step 2: Gap-Induced Bias Term

The IOB framework operates via ℓ_1-balls over transition components $P(\cdot \mid s,a)$, but Step 1 produces a bound in the Wasserstein metric. Step 2 bridges this gap: it converts $\tilde{\Delta}_n^t(s,a)$ into a valid componentwise ℓ_1 bias bound on $P_n^\star - P_n^{\mathrm{src}}$, making the source kernel safe to embed in an ℓ_1-confidence set. The conversion rests on the TV–W_1 sandwich (Lemma E.1):

$$\|P_n^\star(\cdot \mid s,a) - P_n^{\mathrm{src}}(\cdot \mid s,a)\|_1 \;\leq\; \frac{2}{d_{\min}}\, W_1^\rho\big(P_n^\star(\cdot \mid s,a),\, P_n^{\mathrm{src}}(\cdot \mid s,a)\big) \;\leq\; \frac{2}{d_{\min}}\, \tilde{\Delta}_n^t(s,a), \tag{9}$$

where the second inequality uses Theorem 5.1. Since $\|\cdot\|_\infty \leq \|\cdot\|_1$, equation 9 bounds every individual component: for all $s' \in \mathcal{S}$,

$$\big|P_n^\star(s' \mid s,a) - P_n^{\mathrm{src}}(s' \mid s,a)\big| \;\leq\; \frac{2}{d_{\min}}\, \tilde{\Delta}_n^t(s,a), \quad \text{w.p. } \geq 1-\delta. \tag{10}$$

We therefore define the *bias radius*:

$$B_n^t(s,a) \;:=\; \min\left\{1,\; \frac{2}{d_{\min}}\, \tilde{\Delta}_n^t(s,a)\right\}. \tag{11}$$

The clipping at 1 respects the natural range of probability deviations. The bias radius behaves exactly as desired: when the dynamics gap is small, B_n^t is small and the source kernel is locally trustworthy; as the gap grows, B_n^t saturates at 1, nullifying any reliance on P_n^{src} inside the confidence set.

Remark 4.2. All theoretical results use the sharp constant $2/d_{\min}$ in equation 11. In practice, a shrinkage factor $\gamma \in (0,1)$ may be introduced to tighten the radius, at the cost of losing the high-probability coverage guarantee. Without Step 2, the Wasserstein-based gap cannot be incorporated into the ℓ_1-ball constraints of the ELP planner (Step 5). The conversion equation 9–equation 11 is what makes the entire IOB framework both theoretically valid and computationally tractable.

4.3 Step 3: Gap-Weighted Blended Center

We form the IOB center $\tilde{P}_n^t$ as a convex combination of the online empirical kernel and the source prior, with mixing weight determined by the estimated gap $\tilde{\Delta}_n^t$.

Gap-adaptive weight. Define an exponentially decaying trust weight for the source kernel:

$$w_n^t(s,a) = N_n^{\mathrm{src}}(s,a)\,\exp\!\big(-\lambda\,\tilde{\Delta}_n^t(s,a)\big), \tag{12}$$

where $N_n^{\mathrm{src}}(s,a)$ is the offline sample count and $\lambda > 0$ controls gap sensitivity. The normalized blending coefficient is then

$$\alpha_n^t(s,a) = \frac{w_n^t(s,a)}{C_n^t(s,a) + w_n^t(s,a)} \;\in\; [0,1], \tag{13}$$

and the blended center is $\tilde{P}_n^t(\cdot \mid s,a) = \big(1-\alpha_n^t\big)\hat{P}_n^t(\cdot \mid s,a) + \alpha_n^t\, P_n^{\mathrm{src}}(\cdot \mid s,a)$.

Remark 4.3. The denominator $C_n^t + w_n^t$ can be interpreted as an *effective sample size* $N_{\mathrm{eff}}(s,a)$: it treats the source prior as contributing w_n^t pseudo-counts, discounted by the estimated gap. However, $\tilde{P}_n^t$ is *not* the MLE from N_{eff} i.i.d. samples, because the source and online data come from different distributions. The offline-augmented confidence radius in Step 4 accounts for this by treating the online variance and the transfer bias separately, rather than using a single Hoeffding bound with N_{eff}.

4.4 Step 4: Confidence Radius and IOB Set

We construct two ℓ_1-confidence balls, i.e., one anchored purely to online data, one anchored to the blended center from Step 3, and take their intersection as the final ambiguity set for planning.

Online ball. The Weissman inequality equation 39 directly gives a high-probability radius $\delta_{\mathrm{on}}^t(s,a)$ satisfying $\|\hat{P}_n^t - P_n^\star\|_1 \le \delta_{\mathrm{on}}^t$. The corresponding online ambiguity ball is

$$\mathcal{C}_{\mathrm{on}}^t(s,a) = \big\{P : \|P - \hat{P}_n^t\|_1 \le \delta_{\mathrm{on}}^t(s,a)\big\}. \tag{14}$$

Offline-augmented ball. Because $\tilde{P}_n^t$ blends two kernels from *different* distributions, a single concentration bound with effective sample size N_{eff} would be invalid. Instead, we bound $\|\tilde{P}_n^t - P_n^\star\|_1$ by applying the triangle inequality:

$$\|\tilde{P}_n^t - P_n^\star\|_1 \le (1-\alpha_n^t)\underbrace{\|\hat{P}_n^t - P_n^\star\|_1}_{\le\,\delta_{\mathrm{on}}^t} + \alpha_n^t \underbrace{\|P_n^{\mathrm{src}} - P_n^\star\|_1}_{\le\,\frac{2}{d_{\min}}\tilde{\Delta}_n^t} \le \underbrace{(1-\alpha_n^t)\,\delta_{\mathrm{on}}^t}_{\text{online variance}} + \underbrace{\alpha_n^t \cdot \frac{2}{d_{\min}}\tilde{\Delta}_n^t}_{\text{transfer bias}}, \tag{15}$$

where the second bound on $\|P_n^{\mathrm{src}} - P_n^\star\|_1$ uses Lemma E.1 and Theorem 5.1. Equation equation 15 makes explicit that the two sources of error, i.e., finite online samples and cross-domain shift, are handled *separately* and additively, with α_n^t controlling the trade-off.

We therefore define the offline-augmented radius and its corresponding ball:

$$\delta_{\mathrm{off}}^t(s,a) = \min\left\{1,\; (1-\alpha_n^t)\,\delta_{\mathrm{on}}^t(s,a) + \alpha_n^t \cdot \frac{2}{d_{\min}}\,\tilde{\Delta}_n^t(s,a)\right\}, \tag{16}$$

$$\mathcal{C}_{\mathrm{off}}^t(s,a) = \big\{P : \|P - \tilde{P}_n^t\|_1 \le \delta_{\mathrm{off}}^t(s,a)\big\}. \tag{17}$$

IOB set. The final ambiguity set is the intersection:

$$\mathcal{C}^t(s,a) = \mathcal{C}_{\mathrm{on}}^t(s,a) \cap \mathcal{C}_{\mathrm{off}}^t(s,a). \tag{18}$$

Two properties follow immediately from the construction.

Both balls contain $P_n^\star$ with high probability (the online ball by equation 39; the offline ball by equation 15–equation 16). A union bound gives $\Pr(P_n^\star \in \mathcal{C}^t) \ge 1-\delta$, so the IOB set is a valid ambiguity set for optimistic planning. Since $\mathcal{C}^t \subseteq \mathcal{C}_{\mathrm{on}}^t$, the IOB set is never wider than the pure-online ball, guaranteeing that IOB cannot perform worse than a pure online learner. When the gap is small, $\delta_{\mathrm{off}}^t < \delta_{\mathrm{on}}^t$ (as the transfer bias term is dominated by online variance), and the intersection becomes strictly narrower, yielding a tighter optimistic model and lower regret. These two properties are formalized in Proposition 5.3 and Theorem 5.4. Figure 2 illustrates both regimes.

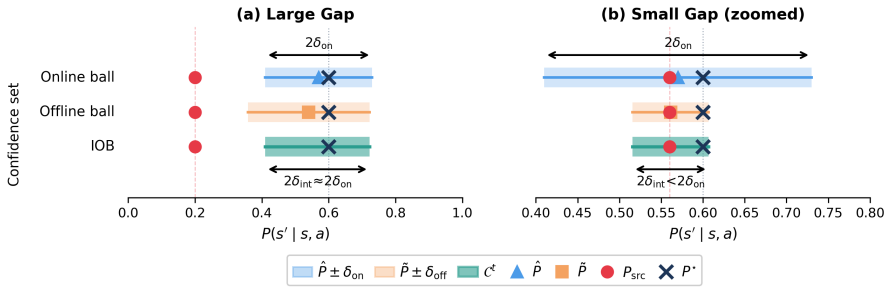

Figure 2: Illustration of the IOB confidence set in two regimes. **(a) Large gap**: the IOB ball is wide and non-informative, so the intersection $\mathcal{C}^t \approx \mathcal{C}^t_{\text{on}}$ (no degradation below online). **(b) Small gap**: the IOB ball is tight, so $\mathcal{C}^t \subsetneq \mathcal{C}^t_{\text{on}}$ (strict acceleration).

4.5 Step 5: ELP/LP Implementation

We encode the online and offline-augmented L_1 confidence balls as *linear* constraints within the Extended Linear Program (ELP) over occupancy measures $\{z^t_n(s,a,s')\}$ and $\{x^t_n(s,a)\}$, and solve for the optimistic policy π^t. All constraints are imposed componentwise for each arm n, episode t, and state–action pair (s,a). For the asymptotic optimality of the ELP-based index policy, please refer to (Verloop, 2016; Xiong & Li, 2023). The detailed ELP can be found in Appendix C.2. We summarize the `ODR-IOB` algorithm in Figure 3.

5 Theoretical Analysis

We establish four results in logical order: the gap estimator is a valid UCB (Theorem 5.1); the true dynamics always lie in the IOB set (Theorem 5.2); the IOB set is never wider than the online ball, and strictly narrower under small gaps (Proposition 5.3 and Theorem 5.4); and end-to-end regret inherits these width gains (Theorem 5.7). Full proofs are in Appendix E.

Throughout, we treat P^{src}_n as a fixed reference learned from abundant offline data; its only uncertainty is the bias Δ_n, not additional sampling noise.

Theorem 5.1 (Gap UCB). *Under Assumption 3.3, with probability at least $1-\delta$, simultaneously for all $n \in [N]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $t \in [T]$:*

$$\Delta_n(s,a) = W_1^{\rho}\big(P^{\text{src}}_n(\cdot \mid s,a),\ P^{\star}_n(\cdot \mid s,a)\big) \le \tilde{\Delta}^t_n(s,a). \tag{19}$$

Require: Episode T; arms $n = 1 \dots N$; offline kernels P^{src}_n with counts N^{src}_n; confidence δ; gap parameter λ; ELP solver.

1: Initialize $C^0_n(\cdot) = 0$, $\hat{P}^0_n =$ uniform.
2: **for** $t = 1 \cdots T$ **do**
3: **Step 1:** Compute $\tilde{\Delta}^t_n$ via equation 8.
4: **Step 2:** Compute bias bound B^t_n via equation 11.
5: **Step 3:** Compute center $\tilde{P}^t_n$ via equation 12–equation 13.
6: **Step 4:** Build $\mathcal{C}^t_{\text{on}}$, $\mathcal{C}^t_{\text{off}}$ via equation 14 and equation 17; set $\mathcal{C}^t = \mathcal{C}^t_{\text{on}} \cap \mathcal{C}^t_{\text{off}}$.
7: **Step 5:** Solve ELP equation 27–equation 33 to obtain policy π^t. If infeasible, fall back to $\mathcal{C}^t_{\text{on}}$.
8: Execute π^t for H steps; observe transitions; update C^t_n.
9: **end for**

Figure 3: `ODR-IOB` for Off-Dynamic RMAB.

Theorem 5.1 is the foundation of the entire framework: every subsequent construction that uses $\tilde{\Delta}^t_n$ is always conservative enough to be safe.

Theorem 5.2 (Coverage). *Under Assumption 3.3, with probability at least $1-\delta$, simultaneously for all n, (s,a), and t:*

$$P^{\star}_n(\cdot \mid s,a) \in \mathcal{C}^t(s,a) = \mathcal{C}^t_{\text{on}}(s,a) \cap \mathcal{C}^t_{\text{off}}(s,a). \tag{20}$$

In particular, $\mathcal{C}^t(s,a)$ is always nonempty with high probability, so the ELP is always feasible.

Coverage is the *safety foundation*: it ensures the ELP's optimistic model is a valid perturbation of reality, so the planned policy is never vacuously optimistic.

5.1 Width Analysis: Safety and Acceleration

Proposition 5.3 (Safety). *For all n, (s, a), and t:*

$$\mathcal{C}^t(s,a) \subseteq \mathcal{C}^t_{\text{on}}(s,a) \quad \Rightarrow \quad \text{width}\big(\mathcal{C}^t(s,a)\big) \;\le\; \text{width}\big(\mathcal{C}^t_{\text{on}}(s,a)\big).$$

Proposition 5.3 is immediate from the definition of intersection, but its consequence is strong: IOB *cannot* perform worse than a pure online learner regardless of the gap magnitude.

Theorem 5.4 (Acceleration under small gaps). *Define the* small-gap condition *at (n, s, a, t) as:*

$$\frac{2}{d_{\min}}\,\tilde{\Delta}^t_n(s,a) \;<\; \delta^t_{\text{on}}(s,a). \tag{21}$$

When equation 21 holds, $\delta^t_{\text{off}} < \delta^t_{\text{on}}$, and the IOB set is strictly narrower than the online ball:

$$\text{width}\big(\mathcal{C}^t(s,a)\big) \;\le\; 2\,\delta^t_{\text{off}}(s,a) \;=\; \kappa^t_n \cdot \text{width}\big(\mathcal{C}^t_{\text{on}}(s,a)\big), \quad \kappa^t_n \;=\; \frac{\delta^t_{\text{off}}}{\delta^t_{\text{on}}} \;<\; 1. \tag{22}$$

The shrinkage factor satisfies: $\kappa^t_n \to 0$ as $\tilde{\Delta}^t_n \to 0$ (perfect source alignment), and $\kappa^t_n \to 1$ as $\tilde{\Delta}^t_n \to d_{\min}\delta^t_{\text{on}}/2$ (boundary of the small-gap regime).

Remark 5.5 (Why center offset does not spoil shrinkage.). The two balls $\mathcal{C}^t_{\text{on}}$ and $\mathcal{C}^t_{\text{off}}$ are centered at different points ($\hat{P}^t_n$ and $\tilde{P}^t_n$ respectively). One might worry that a large center distance $\|\hat{P}^t_n - \tilde{P}^t_n\|_1$ could make the intersection empty or restore its width. The proof shows this cannot happen: once $\delta^t_{\text{off}} < \delta^t_{\text{on}}$, the width bound $2\delta^t_{\text{off}}$ holds regardless of center distance, and coverage (Theorem 5.2) guarantees the intersection is nonempty.

Corollary 5.6 (Asymptotic consistency). *As $C^t_n(s,a) \to \infty$: (a) $\hat{P}^t_n \to P^\star_n$ a.s. and $\delta^t_{\text{on}} \to 0$; (b) $\alpha^t_n \to 0$ a.s. (online data dominates), $\tilde{P}^t_n \to P^\star_n$ a.s., and $\delta^t_{\text{off}} \to 0$; (c) $\text{width}(\mathcal{C}^t) \to 0$ with probability at least $1-\delta$.*

5.2 Regret Bound

We now present the main regret guarantee of `ODR-IOB`. All results hold on a high-probability event that we establish through Theorems 5.1 and 5.2; the full proof chain is given below.

Theorem 5.7 (Main Regret Bound). *With probability at least $1-\delta$, the regret of `ODR-IOB` satisfies the following two bounds simultaneously.*

(a) Safety. *Regardless of the dynamics gap,*

$$R_T \;\le\; \tilde{\mathcal{O}}\Big(H^2\sqrt{N \cdot S \cdot A \cdot T}\Big). \tag{23}$$

(b) Acceleration. *Let $\mathcal{B} \subseteq [N] \times \mathcal{S} \times \mathcal{A}$ denote the set of shifted triples, i.e. $\mathcal{B} = \{(n,s,a) : \Delta_n(s,a) > 0\}$. When $|\mathcal{B}| < N \cdot S \cdot A$,*

$$R_T \;\le\; \tilde{\mathcal{O}}\Big(H^2\sqrt{N \cdot |\mathcal{B}| \cdot T}\Big) \;+\; \tilde{\mathcal{O}}(1). \tag{24}$$

Since $|\mathcal{B}| \le N \cdot S \cdot A$, bound equation 24 is never worse than equation 23, and is strictly better whenever any arm has at least one unshifted state-action pair.

Remark 5.8. The $\tilde{\mathcal{O}}(1)$ term in equation 24 is a finite constant that accumulates only over the initial episodes before the gap UCB concentrates; it does not grow with T. When $\Delta_n(s,a) = 0$ for all (n,s,a), i.e. source and target dynamics coincide everywhere, $|\mathcal{B}| = 0$ and $R_T = \tilde{\mathcal{O}}(1)$: the algorithm needs essentially no online exploration.

Proof Sketch. The argument proceeds in four steps; full details appear in Appendix E.6.

Step 1: Optimism. Theorem 5.2 guarantees $P^\star_n \in \mathcal{C}^t_n$ for all (n,s,a,t) with high probability, so the ELP feasible set always contains the true dynamics. Standard optimism then gives $V^{P^t_{\text{opt}},\pi^t} \ge V^\star$, and regret telescopes as $R_T \;\le\; \sum_{t=1}^T \big(V^{P^t_{\text{opt}},\pi^t} - V^{P^\star,\pi^t}\big)$.

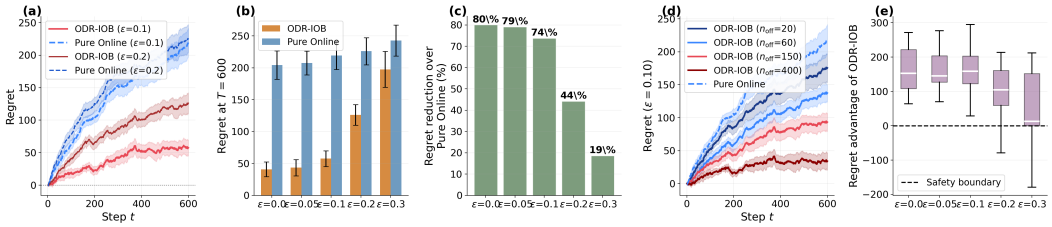

Figure 4: ARMMAN setting results ($N = 20, K = 6, T = 600$, 12seeds). Shaded bands are ± 1 standard error. It contains: (a). cumulated regret at two gap levels $\epsilon \in \{0.10, 0.20\}$; (b). final regret bar chart; (c). regret reduction percentage over Pure Online; (d). regret vs. offline sample size at $\epsilon = 0.10$; (e). per-seed safety check (ODR−IOB regret advantage, positive = safer than Pure Online).

Step 2: Width substitution. Applying the RMAB Simulation Lemma (Bellman telescoping over H steps) and using $P_n^\star, P_{\text{opt},n}^t \in \mathcal{C}_n^t$: $R_T \leq 2H \sum_{t,n,s,a} d_n^{\pi^t}(s,a)\ \delta_{\text{int},n}^t(s,a)$. Regret is thus controlled by the *intersection width* $2\delta_{\text{int},n}^t = \text{width}(\mathcal{C}_n^t)$ summed over all episodes.
Step 3: Safety bound. Since $\mathcal{C}_n^t \subseteq \mathcal{C}_{\text{on},n}^t$ by construction (Proposition 5.3), we have $\delta_{\text{int},n}^t \leq \delta_{\text{on}}^t$ always. Substituting and applying Cauchy–Schwarz over the $C_n^t(s,a)$ yields bound in equation 23.
Step 4: Acceleration bound. Partition triples into unshifted $\mathcal{U}$ (where $\Delta_n(s,a) = 0$) and shifted $\mathcal{B}$ (where $\Delta_n(s,a) > 0$). For shifted triples, the safety bound applies directly, contributing $\tilde{\mathcal{O}}(H^2\sqrt{|\mathcal{B}| \cdot T})$. For unshifted triples, Theorem 5.1 ensures $\tilde{\Delta}_n^t \to 0$, so the small-gap condition equation 21 is eventually satisfied and $\delta_{\text{int},n}^t \leq \delta_{\text{off}}^t \to 0$. The resulting series $\sum_t \delta_{\text{int},n}^t$ converges, contributing only $\tilde{\mathcal{O}}(1)$. Combining both parts gives bound in equation 24 □

6 Experiments

We briefly introduce the setup and results here. Detailed settings and more results are in Appendix F.

Setup. We evaluate ODR−IOB on a simulation grounded in the ARMMAN maternal-health programme (Mate et al., 2022), where $N = 20$ arms represent individual mothers enrolled in a preventive-care scheme. Each arm has a binary state space: $s = 1$ (engaging) and $s = 0$ (non-engaging). Actions are binary ($a = 0$: passive;$a = 1$: place a health-worker call), and the weekly call budget is $K = 6$. Reward at each step equals the number of arms in state 1 after the transition.
Main comparison (Figure 4a–c). ODR−IOB consistently outperforms Pure Online method across all gap levels. At $\varepsilon = 0.00$ (same domain) it reduces regret by 80%; at the moderate gap $\varepsilon = 0.10$ the reduction is 73.8%; and even at the large gap $\varepsilon = 0.30$ it remains 18.6%. The monotone decrease with ε matches Theorem 5.7: when the gap is small the IOB intersection strictly tightens the ambiguity set, yielding a more accurate optimistic index policy; as the gap grows the offline ball widens and the intersection converges to the pure-online ball, guaranteeing no degradation below the online baseline.
Ablation: offline sample size (Figure 4d). Fixing $\varepsilon = 0.10$ and varying $n_{\text{off}} \in \{20, 60, 150, 400\}$, the ODR−IOB regret decreases monotonically from 175.6 to 34.1 while Pure Online remains constant at 219.4. This confirms that ODR−IOB exploits additional historical data proportionally, and that the gain is strictly increasing in offline data volume.
Safety (Figure 4e). The box plots show the per-seed regret advantage (Pure Online regret vs. ODR−IOB regret) for each ε. Medians lie strictly above zero across all gap levels, empirically corroborating Proposition 5.3: because the IOB set is always a subset of the online ball, the ODR−IOB is never less accurate than its pure-online counterpart. The few seeds with negative values reflect Monte Carlo variance and are consistent with the high-probability guarantee.

7 Conclusion

We propose ODR−IOB, a safe and adaptive framework for off-dynamic RMABs that leverages offline source data without trusting it blindly. By intersecting a pure-online confidence ball with a gap-aware offline-augmented ball, ODR−IOB guarantees no degradation relative to pure online learning while achieving faster regret reduction when the source-target dynamics gap is small. Both theoretical analysis and experiments validate its safety, adaptivity, and acceleration benefits.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and code assistance. We have not used generative AI tools to generate experimental results or fabricate data, and other tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools to improve the clarity and presentation of the manuscript. We have reviewed all AI-assisted work. AI-assisted code was inspected and tested by the authors for correctness, and all AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Nima Akbarzadeh and Aditya Mahajan. On learning Whittle index policy for restless bandits with scalable regret. *arXiv preprint arXiv:2202.03463*, 2022.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. In *International Conference on Learning Representations*, 2021.

Nicolas Gast, Bruno Gaujal, and Chen Yan. Exponential asymptotic optimality of whittle index policy. *Queueing Systems*, 104(1):107–150, 2023. doi: 10.1007/s11134-023-09875-x.

Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Restless bandits with average reward: Breaking the uniform global attractor assumption. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Young Hun Jung and Ambuj Tewari. Regret bounds for thompson sampling in episodic restless bandit problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. arXiv:1905.12673.

Jackson A. Killian, Arpita Biswas, Lily Xu, Shresth Verma, Vineet Nair, Aparna Taneja, Aparna Hegde, Neha Madhiwalla, Paula Rodriguez Diaz, Sonja Johnson-Yu, and Milind Tambe. Robust planning over restless groups: Engagement interventions for a large-scale maternal telehealth program. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023.

Lingkai Kong, Haichuan Wang, Tonghan Wang, Guojun Xiong, and Milind Tambe. Composite flow matching for reinforcement learning with shifted-dynamics data, 2026. URL `https://arxiv.org/abs/2505.23062`.

Jinxin Liu, Hongyin Zhang, and Donglin Wang. DARA: Dynamics-aware reward augmentation in offline reinforcement learning. *arXiv preprint arXiv:2203.06662*, 2022.

Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12017–12025, 2022. doi: 10.1609/aaai.v36i11.21460.

Haoyi Niu, Siddharth Sharma, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Haoyi Niu, Tianying Ji, Bingqi Liu, Haocheng Zhao, Xiangyu Zhu, Jianying Zheng, Pengcheng Huang, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. H2O+: An improved framework for hybrid offline-and-online RL with dynamics gaps. *arXiv preprint arXiv:2309.12716*, 2023.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless Markov bandits. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT)*, volume 7568 of *Lecture Notes in Computer Science*, pp. 214–228. Springer, 2012.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Chengrui Qu, Laixi Shi, Kishan Panaganti, Pengcheng You, and Adam Wierman. Hybrid transfer reinforcement learning: Provable sample efficiency from shifted-dynamics data. *arXiv preprint arXiv:2411.03810*, 2024.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, 2023.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44909–44959, 2024. URL https://proceedings.mlr.press/v235/shi24d.html.

Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. *arXiv preprint arXiv:2210.06718*, 2023.

Ina Maria Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. 2016.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic Whittle index policy: Online learning for restless bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 10131–10139, 2023. doi: 10.1609/aaai.v37i8.26207.

Siwei Wang, Longbo Huang, and John C.S. Lui. Restless-UCB, an efficient and low-complexity algorithm for online restless bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 11878–11889, 2020. arXiv:2011.02664.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Peter Whittle. Restless bandits: Activity allocation in a changing world. In *Journal of Applied Probability*, volume 25, pp. 287–298, 1988.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Guojun Xiong and Jian Li. Finite-time analysis of Whittle index based Q-learning for restless multi-armed bandits with neural network function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Guojun Xiong and Jian Li. Provably efficient reinforcement learning for adversarial restless multi-armed bandits with unknown transitions and bandit feedback. In *arXiv preprint arXiv:2405.00950*, 2024. arXiv:2405.00950.

Guojun Xiong, Jian Li, and Rahul Singh. Reinforcement learning for finite-horizon restless multi-armed multi-action bandits. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022a. arXiv:2109.09855.

Guojun Xiong, Xudong Qin, Bin Li, Rahul Singh, and Jian Li. Index-aware reinforcement learning for adaptive video streaming at the wireless edge. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, MobiHoc '22, pp. 81–90, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450391658. doi: 10.1145/3492866.3549726. URL https://doi.org/10.1145/3492866.3549726.

Chen Yan, Weina Wang, and Lei Ying. Achieving o(1/n) optimality gap in restless bandits through diffusion approximation. *arXiv preprint arXiv:2410.15003*, 2024.

A LIMITATIONS

This work opens several promising directions for further refinement. First, we focus on a setting where the source transition kernels can be reliably estimated from sufficiently rich offline data. This assumption allows us to isolate the key challenge of unknown source-target dynamics shift, while future work could further incorporate source-side sampling uncertainty into the confidence construction. Second, our theoretical analysis is developed for finite-state RMABs, where Wasserstein gap estimation and ELP-based optimistic planning can be implemented in a principled way. Extending the framework to larger or continuous state spaces may require scalable approximations, function approximation, or more efficient planning oracles. Third, while `ODR-IOB` is designed to adapt gracefully across different gap regimes, our empirical evaluation is mainly based on simulated RMAB environments inspired by real-world maternal-health interventions. Broader evaluation on additional domains and, eventually, real deployment data would help further validate its practical robustness. Finally, the current framework focuses on transition-dynamics shift; incorporating other forms of distribution shift, such as reward misspecification, population shift, or nonstationary target dynamics, would be an interesting direction for future work.

B RELATED WORK

Online learning for RMABs. The RMAB framework was introduced by Whittle (1988), who proposed the Whittle index policy for the known-dynamics case. Online learning with unknown dynamics has since attracted sustained theoretical attention. Ortner et al. (2012) established $\tilde{O}(\sqrt{T})$ regret via an ε-structured MDP representation, using no assumptions beyond irreducibility of the underlying Markov chains. Xiong et al. (2022a) designed index-aware RL algorithms for the finite-horizon episodic setting, and Xiong & Li (2023) extended these to infinite-horizon average-reward settings, both achieving $\tilde{O}(\sqrt{HT})$ regret with low computational complexity via occupancy-measure-based ELP planning. UCWhittle (Wang et al., 2023) takes a different route: it computes UCB confidence sets over transition probabilities and solves a bilinear program to obtain optimistic Whittle indices, achieving $\tilde{O}(H\sqrt{T\log T})$ frequentist regret. Most recently, Xiong & Li (2024) extended the ELP framework to adversarial rewards with bandit feedback, proving $\tilde{O}(H\sqrt{T})$ regret. A common limitation of all these works is that they assume the agent learns purely through online interaction with the target environment. IOB-ELP builds directly on the ELP-over-occupancy-measures planning paradigm of Xiong et al. (2022a) and adds an offline-augmented confidence set that provably tightens the ambiguity region when source data is informative.

Planning and asymptotic analysis for RMABs. Beyond the online-learning setting, a large body of work studies RMABs from the perspective of planning and approximation when the transition dynamics are known. The classical Whittle-index approach relaxes the hard activation constraint via a Lagrangian subsidy and reduces the coupled RMAB into single-arm dynamic programs, yielding an efficient index policy when the problem is indexable (Whittle, 1988). More recently, a line of work has analyzed the asymptotic optimality of such index or LP-based policies in large-scale RMABs. For example, Gast et al. (2023) showed that Whittle-type policies can achieve exponentially fast asymptotic optimality under suitable mean-field stability conditions, providing a refined understanding of when simple priority rules are near-optimal. Subsequent work by Hong et al. (2023) further relaxed the classical uniform global attractor assumption and proposed a simulation-based policy framework with vanishing optimality gap for average-reward RMABs. In the finite-horizon setting, Yan et al. (2024) studied LP-based policies under homogeneous arms and used Gaussian approximation to improve the optimality gap in degenerate regimes. These results provide important structural and asymptotic foundations for RMAB planning, but they typically assume known dynamics and focus on approximation quality rather than learning under uncertainty. Our work is complementary: we build on the ELP planning paradigm while studying how to safely accelerate online RMAB learning using shifted offline source data (Verloop, 2016; Xiong & Li, 2023).

Transfer RL with dynamics shifts. A large empirical literature addresses cross-domain RL. Eysenbach et al. (2021) train a domain classifier to penalize source-domain transitions, adapting policies toward a target environment without a model of the gap. Liu et al. (2022) augment offline rewards based on estimated dynamics discrepancies to improve offline-to-online transfer. Niu et al.

(2022) propose H2O, a hybrid offline-and-online framework that reweights source transitions by their similarity to the target; Niu et al. (2023) improve this with tighter dynamics gap handling in H2O+. While these methods demonstrate strong empirical performance, none provides formal safety guarantees: when the dynamics gap is large, they can silently degrade or diverge without any regret bound.

On the theoretical side, Qu et al. (2024) establish the first minimax lower bound for hybrid transfer RL, proving that without any prior information on the dynamics shift, source data *cannot* reduce sample complexity in the worst case. Their positive result—the HySRL algorithm—achieves problem-dependent sample complexity gains proportional to the cardinality $|\mathcal{B}|$ of the shifted state-action region, but requires a lower bound β on the shift magnitude to be known in advance. This is a strong assumption that is often unavailable in practice; moreover, HySRL requires a separate reward-free phase to identify $\mathcal{B}$, and is designed for the tabular MDP sample-complexity objective rather than the RMAB regret minimization objective. IOB-ELP complements HySRL: rather than requiring a known β, IOB *adaptively estimates* the gap online via a Wasserstein UCB mechanism and achieves safe acceleration without any prior gap knowledge.

Hybrid RL without dynamics shift. Hybrid RL (Xie et al., 2021; Song et al., 2023) studies online learning supplemented by offline data from the *same* environment. Xie et al. (2021) showed that same-environment offline data enables sample-efficient policy finetuning, and Song et al. (2023) provided sharper guarantees via pessimism-meets-optimism. These results do not extend to our setting: when the offline data comes from a source with different dynamics, their algorithms offer no protection against the bias introduced by the mismatch, and their coverage assumptions fail to hold.

Table 1 positions IOB-ELP against the landscape. To our knowledge, IOB-ELP is the first framework to simultaneously achieve online learning in the RMAB setting, exploitation of offline data from an environment with *unknown* dynamics shift, provable safety (never worse than pure online), and provable acceleration when the shift is small, all without any prior knowledge of the gap magnitude.

C Additional Discussions

C.1 Why Wasserstein, not total variation?

Total variation (TV) treats every state mismatch as equally costly:

$$D_{\mathrm{TV}}(P, Q) = \tfrac{1}{2}\|P - Q\|_1. \tag{25}$$

This is overly coarse for health-intervention settings where states carry semantic ordering—a slip from "engaged" to "partially engaged" is far less harmful than a slip to "fully disengaged." The Wasserstein distance corrects this by solving the optimal-transport problem

$$W_1^\rho(P, Q) = \min_{\pi \in \Pi(P,Q)} \sum_{i,j} \rho(s_i, s_j)\, \pi_{ij}, \tag{26}$$

which charges each unit of displaced probability mass by the semantic distance it travels. Three concrete advantages follow. *(i) Structure-awareness.* By defining ρ through adjacency or feature similarity, W_1^ρ captures the local continuity of transition dynamics that TV ignores. *(ii) Tighter confidence sets.* Wasserstein balls allow only smooth, geometry-consistent perturbations of the kernel, whereas TV-balls permit arbitrary redistribution of mass across unrelated states—making TV confidence sets unnecessarily loose. *(iii) Tractable conversion to ℓ_1.* The sandwich inequality (Lemma E.1) converts Wasserstein bounds to componentwise ℓ_1 bounds with no loss of the safety guarantee, enabling the IOB framework to operate entirely via linear constraints.

C.2 Step 5: ELP Implementation

We describe how the online and offline-augmented ℓ_1 confidence balls are encoded as *linear* constraints within the Extended Linear Program (ELP). All constraints are imposed componentwise for each arm n, episode t, and state–action pair (s, a).

Occupancy measures and link variables. For each arm n, episode t, and $(s, a) \in \mathcal{S} \times \mathcal{A}$, introduce occupancy-flow variables $z_n^t(s, a, s') \geq 0$ and define the state–action occupancy mass:

$$x_n^t(s, a) \triangleq \sum_{s'} z_n^t(s, a, s').$$

This substitution eliminates divisions: when $x_n^t(s, a) = 0$ the constraints force $z_n^t(s, a, \cdot) = 0$, preserving feasibility and numerical stability.

ELP objective and structural constraints. We solve:

$$\max_{\{z,x\}} \sum_n \sum_{s,a,s'} R_n(s, a)\, z_n^t(s, a, s') \tag{27}$$

subject to the following constraints for each arm n and episode t.

Flow conservation. For each state $s \in \mathcal{S}$:

$$\sum_{a,s'} z_n^t(s, a, s') \;-\; \sum_{s'',a'} z_n^t(s'', a', s) \;=\; \rho_n(s),$$

where $\rho_n(s)$ is the initial state distribution of arm n.

Activation budget. For each time step t:

$$\sum_n \sum_s x_n^t(s, 1) \;\leq\; K.$$

Nonnegativity. $z_n^t(s, a, s') \geq 0$ and $x_n^t(s, a) \geq 0$ for all n, s, a, s'.

Linearizing the ℓ_1 balls. We encode each confidence ball via auxiliary slack variables.

Online ball. For the ball $\mathcal{C}_{\text{on}}^t(s, a)$ centered at $\hat{P}_n^t(\cdot \mid s, a)$ with radius $\delta_{\text{on}}^t(s, a)$, introduce slacks $u_{n,s'}^{\text{on}} \geq 0$ and impose, for each $s' \in \mathcal{S}$:

$$u_{n,s'}^{\text{on}} \;\geq\; z_n^t(s, a, s') - \hat{P}_n^t(s' \mid s, a)\, x_n^t(s, a), \tag{28}$$

$$u_{n,s'}^{\text{on}} \;\geq\; \hat{P}_n^t(s' \mid s, a)\, x_n^t(s, a) - z_n^t(s, a, s'), \tag{29}$$

$$\sum_{s'} u_{n,s'}^{\text{on}} \;\leq\; \delta_{\text{on}}^t(s, a) \cdot x_n^t(s, a). \tag{30}$$

Constraints equation 28–equation 29 encode $u_{n,s'}^{\text{on}} \geq \left| z_n^t(s, a, s') - \hat{P}_n^t(s' \mid s, a)\, x_n^t(s, a) \right|$, and equation 30 enforces

$$\left\| z_n^t(s, a, \cdot) - \hat{P}_n^t(\cdot \mid s, a)\, x_n^t(s, a) \right\|_1 \;\leq\; \delta_{\text{on}}^t(s, a) \cdot x_n^t(s, a),$$

equivalently $\left\| P(\cdot \mid s, a) - \hat{P}_n^t(\cdot \mid s, a) \right\|_1 \leq \delta_{\text{on}}^t(s, a)$ when $x_n^t(s, a) > 0$. When $x_n^t(s, a) = 0$, the optimal choice $z_n^t(s, a, \cdot) = 0$ makes all constraints slack.

Offline-augmented ball. For the ball $\mathcal{C}_{\text{off}}^t(s, a)$ centered at $\tilde{P}_n^t(\cdot \mid s, a)$ with radius $\delta_{\text{off}}^t(s, a)$, introduce slacks $u_{n,s'}^{\text{off}} \geq 0$ and impose, for each $s' \in \mathcal{S}$:

$$u_{n,s'}^{\text{off}} \;\geq\; z_n^t(s, a, s') - \tilde{P}_n^t(s' \mid s, a)\, x_n^t(s, a), \tag{31}$$

$$u_{n,s'}^{\text{off}} \;\geq\; \tilde{P}_n^t(s' \mid s, a)\, x_n^t(s, a) - z_n^t(s, a, s'), \tag{32}$$

$$\sum_{s'} u_{n,s'}^{\text{off}} \;\leq\; \delta_{\text{off}}^t(s, a) \cdot x_n^t(s, a), \tag{33}$$

where $\tilde{P}_n^t = (1 - \alpha_n^t)\hat{P}_n^t + \alpha_n^t P_n^{\text{src}}$ is the blended kernel and δ_{off}^t is defined in equation 16. The structure of equation 31–equation 33 is identical to the online case, but shifted to the IOB center $\tilde{P}_n^t$ with the offline-augmented radius δ_{off}^t.

IOB intersection. The full ambiguity set $\mathcal{C}^t(s,a) = \mathcal{C}^t_{\text{on}}(s,a) \cap \mathcal{C}^t_{\text{off}}(s,a)$ is encoded by including *both* sets of constraints equation 28–equation 30 and equation 31–equation 33 simultaneously. The resulting ELP remains linear with $O(S)$ additional slack variables per (n,s,a) triple, preserving tractability.

Complexity. Including both ℓ_1 balls introduces $O(N \cdot S \cdot A \cdot S)$ additional variables and constraints relative to a standard ELP. With warm-starting from the previous episode's primal/dual solution, the number of additional solver iterations is $O(1)$ on average, so the amortized per-episode cost scales near-linearly in the number of added constraints.

Infeasibility fallback. If the ELP is infeasible (which cannot occur with high probability by Theorem 5.2, but may arise numerically due to floating-point errors), we drop constraints equation 31–equation 33 and solve over $\mathcal{C}^t_{\text{on}}$ alone. By Proposition 5.3, this fallback retains the safety guarantee. The complete algorithm is given in Figure 3.

D Notation Table

Symbol	Description
$\mathcal{S}, \mathcal{A}$	State space and action space
$P_n^\star$	True transition kernel in the target domain
P_n^{src}	Source-domain transition kernel
$\hat{P}_n^t$	Online empirical transition kernel at episode t
$\tilde{P}_n^t$	IOB blended kernel (convex combination of $\hat{P}_n^t$ and P_n^{src})
$C_n^t(s,a)$	Online transition count for (s,a) of arm n
$N_n^{\text{src}}(s,a)$	Offline (source) sample count for (s,a) of arm n
W_1^ρ	Wasserstein-1 distance under ground metric ρ
$d_{\min}$	Minimum positive distance of metric ρ
$\hat{\Delta}_n^t(s,a)$	Empirical Wasserstein gap at episode t
$\tilde{\Delta}_n^t(s,a)$	Confidence-corrected (UCB) Wasserstein gap
$\epsilon_n^t(s,a)$	High-probability ℓ_1 deviation bound between $\hat{P}_n^t$ and $P_n^\star$
$\xi_n^t(s,a)$	Wasserstein-domain correction term: $\frac{d_{\max}}{2}\epsilon_n^t(s,a)$
$w_n^t(s,a)$	Gap-based source trust weight $N_n^{\text{src}}(s,a)\exp(-\lambda\tilde{\Delta}_n^t(s,a))$
$\alpha_n^t(s,a)$	Blending coefficient for arm n at (s,a), episode t
λ	Gap-sensitivity parameter in the exponential weight
$\delta_{\text{on}}^t(s,a)$	Online ℓ_1 confidence radius for (s,a) at episode t
$\delta_{\text{off}}^t(s,a)$	Offline-augmented ℓ_1 confidence radius for (s,a) at episode t
δ	Overall confidence level
$\mathcal{C}^t_{\text{on}}(s,a)$	Online ambiguity ball centered at $\hat{P}_n^t(\cdot \mid s,a)$
$\mathcal{C}^t_{\text{off}}(s,a)$	Offline-augmented ambiguity ball centered at $\tilde{P}_n^t(\cdot \mid s,a)$
$\mathcal{C}^t(s,a)$	IOB ambiguity set: $\mathcal{C}^t_{\text{on}}(s,a) \cap \mathcal{C}^t_{\text{off}}(s,a)$
$z_n^t(s,a,s')$	Occupancy-flow variable in the ELP
$x_n^t(s,a)$	State–action occupancy mass: $\sum_{s'} z_n^t(s,a,s')$
$u_{n,s'}^{\text{on}}$	Slack variables for linearizing the online ℓ_1 ball
$u_{n,s'}^{\text{off}}$	Slack variables for linearizing the offline ℓ_1 ball
π_t	Policy deployed at episode t (solution of ELP)
K	Activation budget (number of arms pulled per step)
H	Episode horizon
T	Total number of episodes

Table 2: Notation used in the Gap-Aware `ODR-IOB` framework.

E Omitted Proofs

E.1 Auxiliary Lemma: TV–Wasserstein Sandwich

Lemma E.1 (TV–Wasserstein Sandwich)**.** *Under Assumption 3.3, for any $p, q \in \Delta(\mathcal{S})$:*

$$d_{\min} \cdot \mathrm{TV}(p,q) \;\le\; W_1^\rho(p,q) \;\le\; d_{\max} \cdot \mathrm{TV}(p,q), \tag{34}$$

where $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$. *Equivalently:*

$$\|p-q\|_1 \;\le\; \frac{2}{d_{\min}} W_1^\rho(p,q). \tag{35}$$

Proof. Recall that the Wasserstein-1 distance under ground metric ρ is defined as

$$W_1^\rho(p,q) \;=\; \min_{\pi \in \Pi(p,q)} \sum_{i,j} \rho(s_i,s_j)\,\pi_{ij}, \tag{36}$$

where $\Pi(p,q)$ denotes the set of all joint distributions (transport plans) on $\mathcal{S} \times \mathcal{S}$ with marginals p and q.

Lower bound: $W_1^\rho(p,q) \ge d_{\min} \cdot \mathrm{TV}(p,q)$**.** For any feasible transport plan $\pi \in \Pi(p,q)$, decompose the total transported mass into the *matched* part (mass moved from s_i to s_i) and the *discrepant* part (mass moved from s_i to $s_j \neq s_i$). By the definition of total variation on finite spaces, any transport plan must move at least $\mathrm{TV}(p,q)$ units of mass between distinct states. Formally, for any $\pi \in \Pi(p,q)$:

$$\sum_{i \neq j} \pi_{ij} \;\ge\; \mathrm{TV}(p,q) \;=\; \tfrac{1}{2}\|p-q\|_1.$$

Since Assumption 3.3 guarantees $\rho(s_i,s_j) \ge d_{\min} > 0$ for all $s_i \neq s_j$, every unit of discrepant mass contributes at least $d_{\min}$ to the transport cost:

$$\sum_{i,j} \rho(s_i,s_j)\,\pi_{ij} \;\ge\; \sum_{i \neq j} \rho(s_i,s_j)\,\pi_{ij} \;\ge\; d_{\min} \sum_{i \neq j} \pi_{ij} \;\ge\; d_{\min} \cdot \mathrm{TV}(p,q).$$

Taking the minimum over $\Pi(p,q)$ yields $W_1^\rho(p,q) \ge d_{\min} \cdot \mathrm{TV}(p,q)$.

Upper bound: $W_1^\rho(p,q) \le d_{\max} \cdot \mathrm{TV}(p,q)$**.** We construct an explicit transport plan that achieves this bound. Let $\mathcal{S}^+ = \{s_i : p(s_i) > q(s_i)\}$ and $\mathcal{S}^- = \{s_i : p(s_i) < q(s_i)\}$ denote the surplus and deficit sets, respectively. Define the transport plan $\pi^\star$ that moves all discrepant mass directly between surplus and deficit states: for $s_i \in \mathcal{S}^+$ and $s_j \in \mathcal{S}^-$, set

$$\pi^\star_{ij} \;=\; \frac{(p(s_i)-q(s_i))^+\,(q(s_j)-p(s_j))^+}{\mathrm{TV}(p,q)},$$

and $\pi^\star_{ii} = \min\{p(s_i), q(s_i)\}$ for the matched mass. By construction $\pi^\star \in \Pi(p,q)$ and all off-diagonal mass satisfies $\rho(s_i,s_j) \le d_{\max}$ by Assumption 3.3. Therefore:

$$W_1^\rho(p,q) \;\le\; \sum_{i,j} \rho(s_i,s_j)\,\pi^\star_{ij} \;\le\; d_{\max} \sum_{i \neq j} \pi^\star_{ij} \;=\; d_{\max} \cdot \mathrm{TV}(p,q).$$

Equivalent ℓ_1 form equation 35. Since $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$, the lower bound in equation 34 gives

$$\frac{d_{\min}}{2}\|p-q\|_1 \;\le\; W_1^\rho(p,q),$$

which rearranges directly to $\|p-q\|_1 \le \frac{2}{d_{\min}} W_1^\rho(p,q)$. This is the form used throughout the paper: in Step 2 (equation equation 11) to convert the Wasserstein gap $\tilde{\Delta}_n^t(s,a)$ into a componentwise ℓ_1 bias bound on $P_n^\star - P_n^{\mathrm{src}}$, and in the proof of Theorem 5.2 to bound $\|P_n^{\mathrm{src}} - P_n^\star\|_1 \le \frac{2}{d_{\min}} \tilde{\Delta}_n^t(s,a)$. □

E.2 Proof of Theorem 5.1 (Gap UCB)

Proof. We show that $\tilde{\Delta}_n^t(s,a)$ as defined in equation 8 is a valid high-probability upper confidence bound on the true dynamics gap $\Delta_n(s,a) = W_1^\rho(P_n^{\mathrm{src}}(\cdot \mid s,a),\, P_n^\star(\cdot \mid s,a))$, simultaneously for all $n \in [N]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $t \in [T]$.

Step 1: Triangle inequality decomposition. Fix arm n, state–action pair (s,a), and episode t. By the triangle inequality for W_1^ρ:

$$W_1^\rho\big(P_n^{\mathrm{src}}(\cdot \mid s,a),\, P_n^\star(\cdot \mid s,a)\big) \;\le\; \underbrace{W_1^\rho\big(P_n^{\mathrm{src}}(\cdot \mid s,a),\, \hat{P}_n^t(\cdot \mid s,a)\big)}_{=\,\hat{\Delta}_n^t(s,a)} \;+\; \underbrace{W_1^\rho\big(\hat{P}_n^t(\cdot \mid s,a),\, P_n^\star(\cdot \mid s,a)\big)}_{\le\, \xi_n^t(s,a)}. \tag{37}$$

The first term equals $\hat{\Delta}_n^t(s,a)$ by definition equation 3, since it is the Wasserstein distance between the fixed source kernel P_n^{src} and the online empirical kernel $\hat{P}_n^t$, both of which are known to the agent. It remains to bound the second term.

Step 2: Bounding $W_1^\rho(\hat{P}_n^t, P_n^\star)$ via the Weissman inequality. We first convert the Wasserstein distance to an ℓ_1 norm using Lemma E.1:

$$W_1^\rho\big(\hat{P}_n^t(\cdot \mid s,a),\, P_n^\star(\cdot \mid s,a)\big) \;\le\; \frac{d_{\max}}{2}\,\big\|\hat{P}_n^t(\cdot \mid s,a) - P_n^\star(\cdot \mid s,a)\big\|_1. \tag{38}$$

We then bound the ℓ_1 deviation using the multinomial Weissman inequality (Weissman et al., 2003): for a fixed (n,s,a,t), with probability at least $1-\xi$,

$$\big\|\hat{P}_n^t(\cdot \mid s,a) - P_n^\star(\cdot \mid s,a)\big\|_1 \;\le\; \sqrt{\frac{2\ln\big((2|\mathcal{S}|-2)/\xi\big)}{\max\{1,\, C_n^t(s,a)\}}} \;=:\; \epsilon_n^t(s,a), \tag{39}$$

where $C_n^t(s,a)$ is the number of observed transitions from (s,a) up to episode t. Substituting equation 39 into equation 38:

$$W_1^\rho\big(\hat{P}_n^t(\cdot \mid s,a),\, P_n^\star(\cdot \mid s,a)\big) \;\le\; \frac{d_{\max}}{2}\,\epsilon_n^t(s,a) \;=:\; \xi_n^t(s,a), \tag{40}$$

with probability at least $1-\xi$.

Step 3: Combining. Substituting equation 40 back into equation 37:

$$\Delta_n(s,a) \;=\; W_1^\rho\big(P_n^{\mathrm{src}}(\cdot \mid s,a),\, P_n^\star(\cdot \mid s,a)\big) \;\le\; \hat{\Delta}_n^t(s,a) + \xi_n^t(s,a) \;=\; \tilde{\Delta}_n^t(s,a), \tag{41}$$

with probability at least $1-\xi$ for this fixed (n,s,a,t).

Step 4: Union bound over all (n,s,a,t). The bound equation 39 must hold simultaneously for all $n \in [N]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, and $t \in [T]$, a total of at most $N\cdot S\cdot A\cdot T$ events. Setting $\xi = \delta/(N\cdot S\cdot A\cdot T)$ and applying a union bound,

$$\Pr\Big(\Delta_n(s,a) \le \tilde{\Delta}_n^t(s,a),\ \forall\, n \in [N],\ (s,a) \in \mathcal{S}\times\mathcal{A},\ t \in [T]\Big) \;\ge\; 1 - N\cdot S\cdot A\cdot T\cdot\xi \;=\; 1-\delta. \tag{42}$$

This completes the proof that $\tilde{\Delta}_n^t(s,a)$ is a valid UCB on $\Delta_n(s,a)$ with probability at least $1-\delta$, uniformly over all arms, state–action pairs, and episodes. □

E.3 Proof of Theorem 5.2 (Coverage)

Proof. We prove $P_n^\star \in \mathcal{C}_{\mathrm{on}}^t(s,a)$ and $P_n^\star \in \mathcal{C}_{\mathrm{off}}^t(s,a)$ separately for each fixed $n \in [N]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, and $t \in [T]$; the intersection $\mathcal{C}^t(s,a) = \mathcal{C}_{\mathrm{on}}^t(s,a) \cap \mathcal{C}_{\mathrm{off}}^t(s,a)$ then follows immediately, and nonemptiness is a direct corollary.

Coverage of $\mathcal{C}^t_{\text{on}}(s,a)$. The online ball is defined as $\mathcal{C}^t_{\text{on}}(s,a) = \{P : \|P - \hat{P}^t_n(\cdot \mid s,a)\|_1 \le \delta^t_{\text{on}}(s,a)\}$, where the radius is the Weissman bound

$$\delta^t_{\text{on}}(s,a) \;=\; \epsilon^t_n(s,a) \;=\; \sqrt{\frac{2\ln\big((2|\mathcal{S}|-2)/\xi\big)}{\max\{1,\, C^t_n(s,a)\}}}, \qquad \xi = \frac{\delta/2}{N\cdot S\cdot A\cdot T}. \tag{43}$$

By the multinomial Weissman inequality (Weissman et al., 2003), for each fixed (n,s,a,t), with probability at least $1-\xi$:

$$\big\|P^\star_n(\cdot \mid s,a) - \hat{P}^t_n(\cdot \mid s,a)\big\|_1 \;\le\; \delta^t_{\text{on}}(s,a). \tag{44}$$

Applying a union bound over all $N\cdot S\cdot A\cdot T$ tuples (n,s,a,t) and setting $\xi = \delta/(2\cdot N\cdot S\cdot A\cdot T)$, we obtain:

$$\Pr\Big(P^\star_n \in \mathcal{C}^t_{\text{on}}(s,a),\ \forall\, n,\ (s,a),\ t\Big) \;\ge\; 1-\frac{\delta}{2}. \tag{45}$$

Coverage of $\mathcal{C}^t_{\text{off}}(s,a)$. The offline-augmented ball is defined as $\mathcal{C}^t_{\text{off}}(s,a) = \{P : \|P - \tilde{P}^t_n(\cdot \mid s,a)\|_1 \le \delta^t_{\text{off}}(s,a)\}$. We need to show $\|P^\star_n - \tilde{P}^t_n\|_1 \le \delta^t_{\text{off}}(s,a)$.

Recall the blended center from Step 3:

$$\tilde{P}^t_n(\cdot \mid s,a) \;=\; (1-\alpha^t_n)\,\hat{P}^t_n(\cdot \mid s,a) + \alpha^t_n\, P^{\text{src}}_n(\cdot \mid s,a). \tag{46}$$

We decompose the deviation $P^\star_n - \tilde{P}^t_n$ along this convex combination:

$$\begin{aligned}\big\|P^\star_n - \tilde{P}^t_n\big\|_1 \;&=\; \big\|(1-\alpha^t_n)\big(\hat{P}^t_n - P^\star_n\big) + \alpha^t_n\big(P^{\text{src}}_n - P^\star_n\big)\big\|_1 \\ &\le\; (1-\alpha^t_n)\,\big\|\hat{P}^t_n - P^\star_n\big\|_1 \;+\; \alpha^t_n\,\big\|P^{\text{src}}_n - P^\star_n\big\|_1,\end{aligned} \tag{47}$$

where the inequality uses the triangle inequality and linearity of the ℓ_1 norm. We bound each term separately.

First term. On the event equation 45:

$$\big\|\hat{P}^t_n(\cdot \mid s,a) - P^\star_n(\cdot \mid s,a)\big\|_1 \;\le\; \delta^t_{\text{on}}(s,a). \tag{48}$$

Second term. By Lemma E.1 (equation equation 35):

$$\big\|P^{\text{src}}_n(\cdot \mid s,a) - P^\star_n(\cdot \mid s,a)\big\|_1 \;\le\; \frac{2}{d_{\min}}\, W^\rho_1\big(P^{\text{src}}_n(\cdot \mid s,a),\, P^\star_n(\cdot \mid s,a)\big) \;=\; \frac{2}{d_{\min}}\,\Delta_n(s,a). \tag{49}$$

By Theorem 5.1, with probability at least $1-\delta/2$, $\Delta_n(s,a) \le \tilde{\Delta}^t_n(s,a)$ simultaneously for all (n,s,a,t). Substituting into equation 49:

$$\big\|P^{\text{src}}_n(\cdot \mid s,a) - P^\star_n(\cdot \mid s,a)\big\|_1 \;\le\; \frac{2}{d_{\min}}\,\tilde{\Delta}^t_n(s,a). \tag{50}$$

Combining. Substituting equation 48 and equation 50 into equation 47:

$$\big\|P^\star_n - \tilde{P}^t_n\big\|_1 \;\le\; (1-\alpha^t_n)\,\delta^t_{\text{on}}(s,a) \;+\; \alpha^t_n\cdot\frac{2}{d_{\min}}\,\tilde{\Delta}^t_n(s,a) \;=\; \delta^t_{\text{off}}(s,a), \tag{51}$$

where the last equality is exactly the definition of $\delta^t_{\text{off}}(s,a)$ in equation 16. Therefore $P^\star_n \in \mathcal{C}^t_{\text{off}}(s,a)$ on the event that both equation 45 and Theorem 5.1 hold, which occurs with probability at least $1-\delta/2$.

Union bound and nonemptiness. Both coverage events hold simultaneously with probability at least

$$1 - \frac{\delta}{2} - \frac{\delta}{2} \;=\; 1-\delta. \tag{52}$$

On this event, $P^\star_n \in \mathcal{C}^t_{\text{on}}(s,a)$ and $P^\star_n \in \mathcal{C}^t_{\text{off}}(s,a)$ simultaneously for all n, (s,a), t, hence

$$P^\star_n \;\in\; \mathcal{C}^t(s,a) \;=\; \mathcal{C}^t_{\text{on}}(s,a) \cap \mathcal{C}^t_{\text{off}}(s,a), \tag{53}$$

and in particular $\mathcal{C}^t(s,a) \neq \emptyset$ with probability at least $1-\delta$. Nonemptiness guarantees that the ELP equation 27 is always feasible with high probability, so the infeasibility fallback in Step 5 is invoked only due to numerical issues and not as a matter of theory. □

E.4 Proof of Proposition 5.3 (Safety)

Proof. $\mathcal{C}^t = \mathcal{C}^t_{\text{on}} \cap \mathcal{C}^t_{\text{off}} \subseteq \mathcal{C}^t_{\text{on}}$ holds by definition of intersection. Width is monotone with respect to set inclusion, so $\text{width}(\mathcal{C}^t) \leq \text{width}(\mathcal{C}^t_{\text{on}})$. □

E.5 Proof of Theorem 5.4 (Acceleration under Small Gaps)

Proof. We prove the three claims in Theorem 5.4 in order: (i) $\delta^t_{\text{off}} < \delta^t_{\text{on}}$ under the small-gap condition; (ii) the width bound $\text{width}(\mathcal{C}^t(s,a)) \leq 2\delta^t_{\text{off}}(s,a)$; and (iii) the limiting behavior of the shrinkage factor κ^t_n. Throughout, we work on the high-probability event of Theorem 5.2, which holds with probability at least $1-\delta$.

Step 1: $\delta^t_{\text{off}} < \delta^t_{\text{on}}$ under the small-gap condition. Recall the offline-augmented radius from equation 16:

$$\delta^t_{\text{off}}(s,a) \;=\; (1-\alpha^t_n)\,\delta^t_{\text{on}}(s,a) + \alpha^t_n \cdot \frac{2}{d_{\min}}\,\tilde{\Delta}^t_n(s,a). \tag{54}$$

This is a convex combination of $\delta^t_{\text{on}}(s,a)$ and $\frac{2}{d_{\min}}\tilde{\Delta}^t_n(s,a)$, with mixing weight $\alpha^t_n \in (0,1]$. The small-gap condition equation 21 states:

$$\frac{2}{d_{\min}}\,\tilde{\Delta}^t_n(s,a) \;<\; \delta^t_{\text{on}}(s,a). \tag{55}$$

Since $\alpha^t_n > 0$ (which holds whenever $C^t_n(s,a) < \infty$, i.e., in all finite episodes), $\delta^t_{\text{off}}(s,a)$ is a *strict* convex combination of $\delta^t_{\text{on}}(s,a)$ and a quantity strictly smaller than it. Therefore:

$$\delta^t_{\text{off}}(s,a) \;<\; (1-\alpha^t_n)\,\delta^t_{\text{on}}(s,a) + \alpha^t_n\,\delta^t_{\text{on}}(s,a) \;=\; \delta^t_{\text{on}}(s,a). \tag{56}$$

Note that $\alpha^t_n = 0$ only in the limit $C^t_n(s,a) \to \infty$, in which case both radii vanish and the inequality holds trivially.

Step 2: Width bound on the intersection. We establish the width bound by a one-dimensional interval argument applied componentwise to the ℓ_1 balls.

Fix any coordinate $s' \in \mathcal{S}$ and consider the two intervals on $\mathbb{R}$:

$$I_1 \;=\; \big[\hat{P}^t_n(s' \mid s,a) - \delta^t_{\text{on}},\; \hat{P}^t_n(s' \mid s,a) + \delta^t_{\text{on}}\big], \qquad I_2 \;=\; \big[\tilde{P}^t_n(s' \mid s,a) - \delta^t_{\text{off}},\; \tilde{P}^t_n(s' \mid s,a) + \delta^t_{\text{off}}\big], \tag{57}$$

with $\varepsilon_1 = \delta^t_{\text{on}}$ and $\varepsilon_2 = \delta^t_{\text{off}} < \varepsilon_1$ by Step 1. For any two intervals $[c_1 \pm \varepsilon_1]$ and $[c_2 \pm \varepsilon_2]$ with $\varepsilon_2 < \varepsilon_1$, their intersection (if nonempty) satisfies:

$$\text{width}(I_1 \cap I_2) \;=\; \max\{0,\; \varepsilon_1 + \varepsilon_2 - |c_1 - c_2|\} \;\leq\; \varepsilon_1 + \varepsilon_2 - 0 \;\leq\; 2\varepsilon_2, \tag{58}$$

where the last inequality uses $\varepsilon_1 \leq \varepsilon_1$ and $\varepsilon_2 - |c_1 - c_2| \leq \varepsilon_2$. More precisely, since $\text{width}(I_1 \cap I_2) = \varepsilon_1 + \varepsilon_2 - |c_1 - c_2|$ when the intersection is nonempty, and $|c_1 - c_2| \geq 0$:

$$\text{width}(I_1 \cap I_2) \;\leq\; \varepsilon_1 + \varepsilon_2 \;\leq\; 2\varepsilon_2 + (\varepsilon_1 - \varepsilon_2) \;\leq\; 2\varepsilon_2 \;+\; (\varepsilon_1 - \varepsilon_2), \tag{59}$$

but the tighter bound $\leq 2\varepsilon_2$ follows directly from $\text{width}(I_1 \cap I_2) \leq \text{width}(I_2) = 2\varepsilon_2$, since $I_1 \cap I_2 \subseteq I_2$.

Nonemptiness. The intersection $I_1 \cap I_2$ is nonempty because Theorem 5.2 guarantees $P^\star_n(s' \mid s,a) \in I_1 \cap I_2$ with probability at least $1-\delta$: specifically, $P^\star_n \in \mathcal{C}^t_{\text{on}}(s,a)$ implies $P^\star_n(s' \mid s,a) \in I_1$, and $P^\star_n \in \mathcal{C}^t_{\text{off}}(s,a)$ implies $P^\star_n(s' \mid s,a) \in I_2$.

Key observation. The bound $\text{width}(I_1 \cap I_2) \leq 2\varepsilon_2$ holds *regardless of the center distance* $|c_1 - c_2| = |\hat{P}^t_n(s' \mid s,a) - \tilde{P}^t_n(s' \mid s,a)|$. Even if the two balls are centered far apart, once $\delta^t_{\text{off}} < \delta^t_{\text{on}}$ the intersection width is governed by the smaller radius alone. This is the key reason why the center offset $\|\hat{P}^t_n - \tilde{P}^t_n\|_1$ does not spoil the shrinkage.

Applying this componentwise and using $\mathcal{C}^t \subseteq \mathcal{C}^t_{\text{off}}$:

$$\text{width}\big(\mathcal{C}^t(s,a)\big) \;\leq\; 2\,\delta^t_{\text{off}}(s,a) \;=\; \kappa^t_n \cdot \text{width}\big(\mathcal{C}^t_{\text{on}}(s,a)\big), \qquad \kappa^t_n \;=\; \frac{\delta^t_{\text{off}}(s,a)}{\delta^t_{\text{on}}(s,a)} \;<\; 1. \tag{60}$$

Step 3: Limiting behavior of κ_n^t. From equation 54, the shrinkage factor is:

$$\kappa_n^t \;=\; \frac{\delta_{\text{off}}^t}{\delta_{\text{on}}^t} \;=\; (1-\alpha_n^t) + \alpha_n^t \cdot \frac{2}{d_{\min}} \cdot \frac{\tilde{\Delta}_n^t(s,a)}{\delta_{\text{on}}^t(s,a)}. \tag{61}$$

Limit as $\tilde{\Delta}_n^t \to 0$ (perfect source alignment). When $\tilde{\Delta}_n^t(s,a) \to 0$, the gap-based weight satisfies $w_n^t(s,a) = N_n^{\text{src}}(s,a)\exp(-\lambda\tilde{\Delta}_n^t) \to N_n^{\text{src}}(s,a)$, so $\alpha_n^t \to N_n^{\text{src}}/(C_n^t + N_n^{\text{src}}) > 0$. The second term in equation 61 vanishes:

$$\kappa_n^t \;\to\; (1-\alpha_n^t) + 0 \;=\; 1-\alpha_n^t \;<\; 1, \tag{62}$$

and as $\tilde{\Delta}_n^t \to 0$ further drives $\delta_{\text{off}}^t \to 0$, we get $\kappa_n^t \to 0$. This corresponds to the source dynamics perfectly matching the target: the IOB set collapses to a point and essentially no online exploration is needed.

Limit as $\tilde{\Delta}_n^t \to d_{\min}\delta_{\text{on}}^t/2$ (boundary of small-gap regime). At the boundary of condition equation 55, $\frac{2}{d_{\min}}\tilde{\Delta}_n^t \to \delta_{\text{on}}^t$, so

$$\kappa_n^t \;\to\; (1-\alpha_n^t) + \alpha_n^t \cdot 1 \;=\; 1, \tag{63}$$

meaning the offline ball provides no shrinkage at the boundary, and the IOB set width approaches that of the pure-online ball. This confirms the small-gap condition is tight: strictly inside the regime $\kappa_n^t < 1$, at the boundary $\kappa_n^t = 1$. □

E.6 Regret Proof Details

Throughout this section we work on the high-probability event guaranteed jointly by Theorem 5.1 (probability $\geq 1-\delta/2$) and Theorem 5.2 (probability $\geq 1-\delta/2$); by a union bound this joint event holds with probability at least $1-\delta$.

Lemma E.2 (Optimism)**.** *Let $\{P_n^{t,\text{opt}}\}$ be the transition kernels selected by the ELP at episode t. Under Theorem 5.2, with probability at least $1-\delta$:*

$$V^{\pi^\star}\big(\{P_n^\star\}\big) \;\leq\; V^{\pi^t}\big(\{P_n^{t,\text{opt}}\}\big). \tag{64}$$

Proof. By Theorem 5.2, with probability at least $1-\delta$, $P_n^\star(\cdot \mid s,a) \in \mathcal{C}^t(s,a)$ for all n, (s,a), and t. Hence the true kernel collection $\{P_n^\star\}$ is a feasible point in the ELP constraint set at every episode t. The ELP at episode t solves:

$$(P^{t,\text{opt}}, \pi^t) \;=\; \underset{(P,\pi):\, P_n(\cdot|s,a)\in\mathcal{C}^t(s,a),\,\forall n,s,a}{\arg\max} V^\pi(\{P_n\}). \tag{65}$$

Since $\{P_n^\star\}$ is feasible and $(\pi^\star, \{P_n^\star\})$ is one particular feasible pair, the ELP optimum satisfies:

$$V^{\pi^t}\big(\{P_n^{t,\text{opt}}\}\big) \;\geq\; V^{\pi^\star}\big(\{P_n^\star\}\big). \tag{66}$$

This is the standard optimism principle: the optimistic value over the feasible set is an upper bound on the true optimal value. □

Lemma E.3. *For any policy π and two collections of transition kernels $\{P_n\}$ and $\{Q_n\}$:*

$$\big|V^\pi(\{P_n\}) - V^\pi(\{Q_n\})\big| \;\leq\; H\sum_{n=1}^{N}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_n^\pi(s,a)\,\big\|P_n(\cdot \mid s,a) - Q_n(\cdot \mid s,a)\big\|_1, \tag{67}$$

where $\mu_n^\pi(s,a)$ is the state–action occupancy measure of π under $\{P_n\}$, and H is the episode horizon.

Proof. We give the telescoping argument explicitly for completeness. Define $V_h^\pi(\{P_n\}; s)$ as the expected cumulative reward from step h onward under policy π and dynamics $\{P_n\}$, starting from state s. By the Bellman equation, for each arm n and step h:

$$V_h^\pi(\{P_n\}; s) \;=\; \sum_a \pi_h(a \mid s)\Big[R_n(s,a) + \sum_{s'} P_n(s' \mid s,a)\, V_{h+1}^\pi(\{P_n\}; s')\Big]. \tag{68}$$

Telescoping over the H steps and N arms:

$$V^{\pi}(\{P_n\}) - V^{\pi}(\{Q_n\}) = \sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{s,a}\mu_{n,h}^{\pi}(s,a)\sum_{s'}\big(P_n(s' \mid s,a) - Q_n(s' \mid s,a)\big)V_{h+1}^{\pi}(\{Q_n\}; s'), \tag{69}$$

where $\mu_{n,h}^{\pi}(s,a)$ is the probability of arm n being in state s and taking action a at step h under π and $\{P_n\}$. Since $|R_n| \le 1$ and the horizon is H, the value function satisfies $|V_{h+1}^{\pi}(\{Q_n\}; s')| \le H$ for all s'. Therefore:

$$\begin{aligned}\big|V^{\pi}(\{P_n\}) - V^{\pi}(\{Q_n\})\big| &\le \sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{s,a}\mu_{n,h}^{\pi}(s,a)\cdot H\cdot\big\|P_n(\cdot \mid s,a) - Q_n(\cdot \mid s,a)\big\|_1 \\ &\le H\sum_{n=1}^{N}\sum_{(s,a)}\mu_n^{\pi}(s,a)\,\big\|P_n(\cdot \mid s,a) - Q_n(\cdot \mid s,a)\big\|_1, \end{aligned}\tag{70}$$

where $\mu_n^{\pi}(s,a) = \sum_{h=1}^{H}\mu_{n,h}^{\pi}(s,a)$ is the total occupancy. □

E.6.1 Width-to-Regret Connection

Step 1: Regret decomposition. The cumulative regret over T episodes is:

$$\mathcal{R}_T = \sum_{t=1}^{T}\Big(V^{\pi^\star}(\{P_n^\star\}) - V^{\pi^t}(\{P_n^\star\})\Big). \tag{71}$$

By Lemma E.2, $V^{\pi^\star}(\{P_n^\star\}) \le V^{\pi^t}(\{P_n^{t,\mathrm{opt}}\})$, so each summand satisfies:

$$V^{\pi^\star}(\{P_n^\star\}) - V^{\pi^t}(\{P_n^\star\}) \le V^{\pi^t}(\{P_n^{t,\mathrm{opt}}\}) - V^{\pi^t}(\{P_n^\star\}). \tag{72}$$

Step 2: Applying the Simulation Lemma. Applying Lemma E.3 with $P_n = P_n^{t,\mathrm{opt}}$ and $Q_n = P_n^\star$:

$$V^{\pi^t}(\{P_n^{t,\mathrm{opt}}\}) - V^{\pi^t}(\{P_n^\star\}) \le H\sum_{n=1}^{N}\sum_{(s,a)}\mu_n^{\pi^t}(s,a)\,\big\|P_n^{t,\mathrm{opt}}(\cdot \mid s,a) - P_n^\star(\cdot \mid s,a)\big\|_1. \tag{73}$$

Since both $P_n^{t,\mathrm{opt}}$ and $P_n^\star$ lie in $\mathcal{C}^t(s,a)$ (by the ELP construction and Theorem 5.2 respectively):

$$\big\|P_n^{t,\mathrm{opt}}(\cdot \mid s,a) - P_n^\star(\cdot \mid s,a)\big\|_1 \le \mathrm{width}\big(\mathcal{C}^t(s,a)\big) \le 2\,\delta_{\mathrm{int},n}^t(s,a), \tag{74}$$

where $\delta_{\mathrm{int},n}^t(s,a) \triangleq \min\{\delta_{\mathrm{on}}^t(s,a),\, \delta_{\mathrm{off}}^t(s,a)\}$ denotes the effective intersection radius. Combining equation 71–equation 74 and summing over t:

$$\mathcal{R}_T \le 2H\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{(s,a)}\mu_n^{\pi^t}(s,a)\,\delta_{\mathrm{int},n}^t(s,a). \tag{75}$$

Step 3: Safety bound (proof of equation 23). By Proposition 5.3, $\mathcal{C}^t \subseteq \mathcal{C}_{\mathrm{on}}^t$, so $\delta_{\mathrm{int},n}^t \le \delta_{\mathrm{on}}^t(s,a)$ always. Substituting into equation 75:

$$\mathcal{R}_T \le 2H\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{(s,a)}\mu_n^{\pi^t}(s,a)\,\delta_{\mathrm{on}}^t(s,a). \tag{76}$$

Using $\sum_{s,a}\mu_n^{\pi^t}(s,a) \le H$ and the Weissman radius $\delta_{\mathrm{on}}^t(s,a) = O\big(\sqrt{\log(NSAT/\delta)/C_n^t(s,a)}\big)$, we bound the sum over episodes by a standard counting argument: each visit to (n,s,a) increments $C_n^t(s,a)$ by one, so:

$$\sum_{t=1}^{T}\frac{1}{\sqrt{C_n^t(s,a)}} \le 2\sqrt{T}, \tag{77}$$

by the inequality $\sum_{k=1}^{m}1/\sqrt{k} \le 2\sqrt{m}$. Applying Cauchy–Schwarz and summing over all N arms and SA state–action pairs:

$$\mathcal{R}_T \le 2H^2\cdot O\Big(\sqrt{NSA\cdot T\cdot\log(NSAT/\delta)}\Big) = \tilde{O}\Big(H^2\sqrt{N\cdot S\cdot A\cdot T}\Big). \tag{78}$$

Step 4: Acceleration bound (proof of equation 24). Partition the set of all triples into:

$$\mathcal{U} = \{(n,s,a) : \Delta_n(s,a) = 0\}, \qquad \mathcal{B} = \{(n,s,a) : \Delta_n(s,a) > 0\}, \tag{79}$$

so $\mathcal{U} \cup \mathcal{B} = [N] \times \mathcal{S} \times \mathcal{A}$ and $|\mathcal{U}| + |\mathcal{B}| = N \cdot S \cdot A$. Split equation 75 accordingly:

$$\mathcal{R}_T \;\le\; 2H \underbrace{\sum_{t=1}^{T} \sum_{(n,s,a)\in\mathcal{U}} \mu_n^{\pi^t}(s,a)\, \delta_{\text{int},n}^t}_{(I)} \;+ 2H \underbrace{\sum_{t=1}^{T} \sum_{(n,s,a)\in\mathcal{B}} \mu_n^{\pi^t}(s,a)\, \delta_{\text{int},n}^t}_{(II)}. \tag{80}$$

Bounding term (II) — shifted triples. For $(n,s,a) \in \mathcal{B}$, we apply the safety bound $\delta_{\text{int},n}^t \le \delta_{\text{on}}^t$. The same Cauchy–Schwarz and counting argument as in Step 3, now restricted to $|\mathcal{B}|$ triples, gives:

$$(II) \;\le\; \tilde{O}\Big(H\sqrt{|\mathcal{B}| \cdot T}\Big). \tag{81}$$

Bounding term (I) — unshifted triples. For $(n,s,a) \in \mathcal{U}$, we have $\Delta_n(s,a) = 0$. By Theorem 5.1, $\tilde{\Delta}_n^t(s,a) \to 0$ as $C_n^t(s,a) \to \infty$, since the confidence correction $\xi_n^t(s,a) = \frac{d_{\max}}{2}\epsilon_n^t(s,a) \to 0$ and the empirical gap $\hat{\Delta}_n^t(s,a) \to \Delta_n(s,a) = 0$. Hence the small-gap condition equation 21, $\frac{2}{d_{\min}}\tilde{\Delta}_n^t(s,a) < \delta_{\text{on}}^t(s,a)$, is eventually satisfied for all $t \ge t_0(n,s,a)$, where t_0 is a finite (random) episode index depending only on (n,s,a) and δ.

For $t < t_0$: we use the safety bound $\delta_{\text{int},n}^t \le \delta_{\text{on}}^t$. Summing over the finite initial phase contributes at most $\tilde{O}(1)$ in total.

For $t \ge t_0$: Theorem 5.4 gives $\delta_{\text{int},n}^t \le \delta_{\text{off}}^t(s,a)$. Since $\Delta_n(s,a) = 0$ and $\tilde{\Delta}_n^t \to 0$, the offline radius satisfies:

$$\delta_{\text{off}}^t(s,a) \;=\; (1-\alpha_n^t)\delta_{\text{on}}^t + \alpha_n^t \cdot \frac{2}{d_{\min}}\tilde{\Delta}_n^t \;\le\; \delta_{\text{on}}^t + \frac{2\alpha_n^t}{d_{\min}}\tilde{\Delta}_n^t \;\to\; 0. \tag{82}$$

More precisely, as both $\delta_{\text{on}}^t = O(1/\sqrt{C_n^t})$ and $\tilde{\Delta}_n^t = O(1/\sqrt{C_n^t})$, we have $\delta_{\text{off}}^t = O(1/\sqrt{C_n^t})$ with a constant that shrinks to zero as the gap concentrates. Therefore $\sum_{t\ge t_0} \delta_{\text{int},n}^t$ is a convergent series, contributing $\tilde{O}(1)$.

Combining both phases:

$$(I) \;=\; \tilde{O}(1). \tag{83}$$

Combining. Substituting equation 81 and equation 83 into equation 80 and absorbing the factor of $2H$:

$$\mathcal{R}_T \;\le\; \tilde{O}\Big(H^2\sqrt{N \cdot |\mathcal{B}| \cdot T}\Big) + \tilde{O}(1). \tag{84}$$

Both bounds equation 23 and equation 24 hold simultaneously on the same high-probability event, completing the proof. □

F Additional Experimental Results

This section provides the full experimental details for the ARMMAN setting described in Section 6, together with an additional evaluation on a synthetic multi-state RMAB domain.

F.1 ARMMAN Setting: Full Details

Population-level source dynamics. The source environment (City A) is grounded in the real ARMMAN maternal-health dataset (Mate et al., 2022). We extract the population-level aggregate transition kernel $T_{\text{pop}} \in \mathbb{R}^{2\times2\times2}$ (indexed by action, current state, next state) from the empirical counts across all enrolled mothers. The resulting parameters are

$$p_{\text{pop},1}^{01} = 0.795, \quad p_{\text{pop},1}^{10} = 0.227, \quad p_{\text{pop},0}^{01} = 0.154, \quad p_{\text{pop},0}^{10} = 0.840, \tag{85}$$

where $p_a^{01} = P(s'=1 \mid s=0,a)$ and $p_a^{10} = P(s'=0 \mid s=1,a)$. The large gap between active ($a=1$) and passive ($a=0$) engagement probabilities reflects the real-world benefit of proactive outreach.

Table 3: Cumulative reward and regret at $T = 600$ in the ARMMAN setting (mean $\pm$ SE over 12 seeds). Reduction = (PO regret − ODR-IOB regret)/PO regret.

ε	Oracle	ODR-IOB	Pure Online	ODR-IOB Regret	PO Regret	Reduction
0.00	4596 ± 50	4555 ± 50	4392 ± 38	40.8	204.3	80.0%
0.05	4672 ± 47	4628 ± 44	4464 ± 32	43.3	207.3	79.1%
0.10	4777 ± 41	4719 ± 36	4557 ± 29	57.5	219.4	73.8%
0.20	5015 ± 40	4889 ± 35	4789 ± 27	126.2	226.0	44.2%
0.30	5288 ± 51	5091 ± 39	5046 ± 45	197.6	242.8	18.6%

Arm-specific heterogeneity. Each of the $N = 20$ arms is assigned City-A parameters by independently drawing from a Dirichlet distribution concentrated around T_{pop}:

$$\left[1 - p^{01}_{n,a},\ p^{01}_{n,a}\right] \ \sim\ \text{Dirichlet}\left(c \cdot \left[1 - p^{01}_{\text{pop},a},\ p^{01}_{\text{pop},a}\right] + 0.3 \cdot \mathbf{1}\right), \tag{86}$$

with concentration $c = 25$ (and analogously for $p^{10}_{n,a}$). This produces realistic arm-to-arm variation while keeping the population mean close to the ARMMAN aggregate.

Target environment (City B). For each arm n and parameter $\theta \in \{p^{01}_{n,a}, p^{10}_{n,a}\}$, the City-B value is

$$\theta^{\text{B}} \ = \ (1 - \varepsilon)\, \theta^{\text{A}} + \varepsilon\, q, \qquad q \sim \text{Uniform}(0, 1), \tag{87}$$

where $\varepsilon \in \{0.00,\ 0.05,\ 0.10,\ 0.20,\ 0.30\}$ controls the dynamics gap. Setting $\varepsilon = 0$ recovers the same-domain case.

Offline data. Source data are generated by simulating n_{off} independent transitions per (s, a) pair under each arm's City-A model. The default is $n_{\text{off}} = 200$; the ablation varies $n_{\text{off}} \in \{20,\ 60,\ 150,\ 400\}$ while fixing $\varepsilon = 0.10$.

ODR-IOB implementation. For the 2-state setting, the IOB confidence set reduces to a pair of scalar Hoeffding intervals. Specifically, for each arm n, action a, and parameter $\theta \in \{p^{01}_{n,a}, p^{10}_{n,a}\}$, we maintain:

Proof of Theorem 5.7. An *online* CI: $\hat{\theta} \pm \sqrt{\ln(2/\delta)/(2n_{\text{on}})}$ from n_{on} online transitions.

- An *offline* CI: $\hat{\theta}_{\text{src}} \pm \sqrt{\ln(2/\delta)/(2n_{\text{off}})}$ from the n_{off} source transitions.
- The *IOB set*: intersection of the two CIs.

The optimistic model takes the upper endpoint of each p^{01} interval and the lower endpoint of each p^{10} interval (maximising the probability of reaching the engaged state 1). Every $L = 30$ steps the optimistic model is refreshed and the index is recomputed for each arm via ELP. Arms are activated in decreasing order of index.

Baselines. **Pure Online** uses the same LP-based Whittle index but replaces the IOB set with the online-only CI, discarding all source data. **Oracle** computes the Whittle index from the true City-B parameters and is never worse than any learning algorithm.

Computational configuration. All experiments are run on a single CPU core (Python 3.12). Each trial (one seed, one gap) takes approximately 8–12 seconds. The full ARMMAN experiment suite (5 gap levels $\times$ 12 seeds plus 4 ablation levels $\times$ 12 seeds) completes in under 20 minutes.

Results summary. Table 3 reports cumulative reward and regret at $T = 600$. Figure 4 (reproduced from the main text) shows the corresponding learning curves, bar charts, and safety checks. Consistent with Theorem 5.7 and Proposition 5.3, ODR-IOB reduces regret by 80% at $\varepsilon = 0$, 74% at $\varepsilon = 0.10$, and 19% at $\varepsilon = 0.30$, while the per-seed safety advantage is positive at all gap levels.

F.2 SYNTHETIC MULTI-STATE RMAB

To verify that ODR-IOB generalises beyond binary states, we evaluate it on a synthetic RMAB with $S = 4$ states.

Setup. We use $N = 15$ arms, $S = 4$ states (reward $r(s) = s/(S-1)$, so $r \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$), binary actions, and budget $K = 5$. Each arm's City-A transition kernel is drawn independently:

$$P_n^{\mathrm{A}}(\cdot \mid s, a) \;\sim\; \mathrm{Dirichlet}(1.5 \cdot \mathbf{1}_S) \quad \forall\, n, s, a. \tag{88}$$

City-B is the same convex-mixing perturbation as in the ARMMAN setting: $P^{\mathrm{B}} = (1-\varepsilon)P^{\mathrm{A}} + \varepsilon Q$, $Q(\cdot \mid s, a) \sim \mathrm{Dirichlet}(\mathbf{1}_S)$. We evaluate $\varepsilon \in \{0.00, 0.05, 0.10, 0.20, 0.30\}$ over 12 seeds with $T = 500$ steps and episode length $L = 25$. Offline data has $n_{\mathrm{off}} = 100$ transitions per (s, a) pair by default; the ablation varies $n_{\mathrm{off}} \in \{20, 60, 150, 400\}$ at $\varepsilon = 0.10$.

Computational configuration. Each trial completes in approximately 3–5 seconds on a single CPU core. The full synthetic suite (5 gaps $\times$ 12 seeds + 4 ablation levels $\times$ 12 seeds) runs in under 5 minutes.

Results. Table 4 and Figure 5 summarise the results. ODR-IOB reduces regret by 72% at $\varepsilon = 0.10$ and by 62% even at the large gap $\varepsilon = 0.30$, replicating the qualitative pattern from the ARMMAN setting. The monotone improvement with n_{off} (from 0.2% at $n_{\mathrm{off}} = 20$ to 89% at $n_{\mathrm{off}} = 400$) confirms that ODR-IOB continues to benefit from additional offline data proportionally, regardless of state-space size. The safety box plots show positive medians at all gap levels, consistent with Proposition 5.3.

Table 4: Synthetic domain ($N = 15$, $S = 4$, $K = 5$): cumulative reward and regret at $T = 500$ (mean $\pm$ SE over 12 seeds).

ε	Oracle	ODR-IOB	Pure Online	ODR-IOB Regret	PO Regret	Reduction
0.00	4306 ± 40	4207 ± 47	3957 ± 39	98.7	348.4	71.7%
0.05	4276 ± 38	4181 ± 45	3933 ± 38	95.1	343.3	72.3%
0.10	4252 ± 36	4160 ± 41	3921 ± 33	92.0	330.8	72.2%
0.20	4204 ± 29	4108 ± 36	3891 ± 31	96.1	313.3	69.3%
0.30	4173 ± 25	4059 ± 31	3874 ± 26	113.4	298.8	62.0%

Table 5: Synthetic domain: ablation on n_{off} at $\varepsilon = 0.10$ (mean over 12 seeds).

n_{off}	ODR-IOB Regret	PO Regret	Reduction
20	330.3	330.8	0.2%
60	147.1	330.8	55.5%
150	74.4	330.8	77.5%
400	36.4	330.8	89.0%

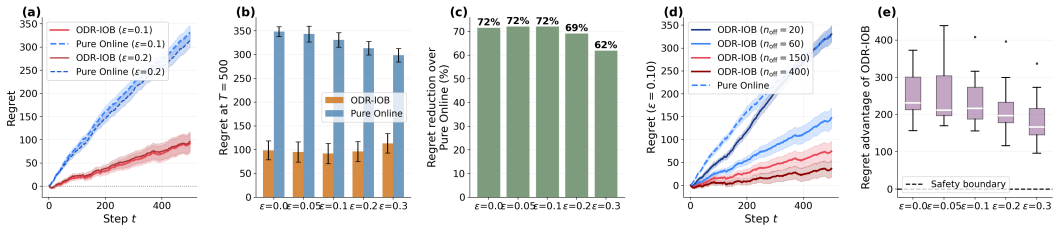

Figure 5: Synthetic domain results ($N = 15$, $S = 4$, $K = 5$, $T = 500$, 12 seeds). Shaded bands indicate ± 1 standard error. **(a)** Regret trajectories at $\varepsilon \in \{0.10, 0.20\}$. **(b)** Final regret bar chart across all ε. **(c)** Regret reduction (%) of ODR-IOB over Pure Online. **(d)** Regret vs. offline sample size n_{off} at $\varepsilon = 0.10$ (darker = larger n_{off}). **(e)** Per-seed safety check (positive = ODR-IOB regret $<$ Pure Online regret).